A STATISTICAL METHOD FOR SYNTACTIC DIALECTOMETRY

Nathan C. Sanders

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Linguistics
Indiana University

October 2010

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements of the degree of Doctor of Philosophy.

Doctoral Committee

Sandra Kübler, Ph.D.
(Principal Advisor)

Markus Dickinson, Ph.D.

Steven Franks, Ph.D.

June 11, 2010

Michael Gasser, Ph.D.

To my parents.

Acknowledgements

First, I would like to thank Sandra Kübler. I could not have asked for a better, more supportive advisor. Sandra provided endless suggestions and help for all aspects of my career as a graduate student.

The rest of my committee deserves thanks as well: Markus Dickinson, Michael Gasser, and Steven Franks all contributed valuable advice and suggestions while I was working on my dissertation, as well as during my earlier work on English.

Other dialect researchers contributed to this dissertation: Henrik Rosenkvist graciously gave me access to his current version of Swediasyn. That data forms the basis of my work here; without it, I would not have a dissertation. Therese Leinonen donated her high-resolution outline maps of Sweden and Swedish-speaking Finland. Her hard work made my results look much better than they would have otherwise.

I also want to thank the members of the Parsing Reading Group, both for putting up with my crazy ideas and often suggesting equally crazy ones in return. They helped me solve a number of practical problems with the dissertation. Stephanie Dickinson was also a great help with statistical tests. I recommend the services of the IU Statistical Consulting Center to every graduate student at IU.

I want to thank the members of the IU Aikido Club and the North Central church for helping me keep my body and spirit healthy even while I was busy exercising my mind. Specifically, Fred and Maryrose offered to feed me far more often than I deserved, Guy Haskell gave advice based on his

own path through graduate school, and Dan Speer, Nathan Mishler, Don Cross, and Sungkyoung Lee provided a lot of moral support.

Finally, the Indiana University motto is “Lux et veritas”, Latin for “Light and truth”. I humbly acknowledge that God is the source of all light and truth, and I hope that this dissertation reflects a tiny part of that truth.

Nathan Sanders

A STATISTICAL METHOD FOR SYNTACTIC DIALECTOMETRY

This dissertation establishes the utility and reliability of a statistical distance measure for syntactic dialectometry, expanding dialectometry's methods to include syntax as well as phonology and the lexicon. It establishes the measure's reliability by comparing its results to those of dialectology and phonological dialectometry on Swedish dialects, as well as evaluating variant parameter settings. The research questions of this dissertation are (1) whether a statistical measure of syntax for dialectometry will reproduce the results of syntactic dialectology and phonological dialectometry and (2) what parameter settings produce results most similar to dialectology's results.

Statistical dialect distance is defined in two parts: a feature set that captures linguistic properties and a measure of dissimilarity that combines two sites' features into a single number. This dissertation uses feature sets from previous work: trigrams (Nerbonne & Wiersma, 2006) and leaf-ancestor paths (Sanders, 2007). In addition, it introduces two other feature sets: leaf-head paths based on dependencies and phrase-structure rules. This dissertation uses the measure R (Nerbonne & Wiersma 2006) as well as measures from information theory: Kullback-Leibler and Jensen-Shannon divergences and cosine similarity. This statistical distance is tested on the Swediasyn, a corpus of interviews recorded in villages throughout Sweden. After the distance was measured, the distances were processed and then compared with existing dialectology results.

Unlike previous work, significant distances were measured between dialect corpora in this dissertation. When these distances are mapped to the geography of Sweden, they reproduce the traditional dialect regions of Sweden. There is weak correlation with geographic distance, but good agreement between dialectometric syntactic and phonological distance. Comparing specific dialect features with those of dialectology is inconclusive; better comparison methods are needed.

Contents

Acknowledgements **v**

Abstract **vii**

1 Introduction **1**

1.1 Overview of Dialectometry 2

Syntax and Dialectometry 4

1.2 Overview of the Dissertation 6

2 Questions **9**

2.1 Question 1 : Agreement with Dialectology 11

Definition of Dialectology Terms 11

Features 12

Isogloss Boundaries 13

Isogloss Bundles 13

Distances 16

2.2 Question 2 : Variations on the Measure 17

Definition of dialectometry terms . . . 17
Feature Sets . . . 19
Distance Measures . . . 20
2.3 Question 3 : Agreement with Phonological Dialectometry . . . 21

3 Methods **22**

3.1 Related Work . . . 22
Séguy . . . 22
Goebl . . . 24
3.2 Dialectometry . . . 27
Syntactic Distance . . . 28
Syntax Features . . . 34
Alternate Feature Sets . . . 38
Combining Feature Sets . . . 42
Alternate Distance Measures . . . 42
3.3 Input Processing . . . 45
SweDiaSyn . . . 46
Talbanken . . . 46
Parsing . . . 47
3.4 Output Analysis . . . 48
Permutation test . . . 48
Cluster Analysis . . . 49
Multi-dimensional scaling . . . 58

Correlation . . . 58
Feature Ranking . . . 60
3.5 Conclusion . . . 63

4 Results **65**

4.1 Parameter Settings . . . 66
4.2 Significant Distances . . . 66
Significance by Measure . . . 69
Significance by Feature Set . . . 70
4.3 Correlation . . . 71
Analysis . . . 75
Inter-measure Correlation . . . 75
Correlation with Corpus Size . . . 76
4.4 Clusters . . . 79
Consensus Trees . . . 84
Composite Cluster Maps . . . 94
4.5 Multi-Dimensional Scaling . . . 98
4.6 Features . . . 102
4.7 wiersma-normalization . . . 103
Trigram Features . . . 106
Trigrams with Overuse Normalization . . . 106
Variation Across Feature Sets . . . 112
Phrase-structure rule features . . . 112
4.8 Conclusion . . . 119

5 Discussion **120**

5.1 Comparison to Syntactic Dialectology 120

General Expectations 121

Dialect Regions 122

Dialect Features 123

Conclusion 137

5.2 Comparison to Phonological Dialectometry 137

5.3 Comparison to Syntactic Dialectometry 142

6 Conclusion **144**

6.1 Future Work 144

6.2 Conclusion 146

List of Tables

3.1 Example dissimilarities 53

3.2 Example dissimilarities 58

4.1 Settings for the five parameters tested 66

4.2 Size of Interview Sites 67

4.3 Number of non-significant distances for sample size 1000, 1 normalization 67

4.4 Number of non-significant distances for complete sites, 1 normalization 68

4.5 Number of non-significant distances for sample size 1000, 5 normalizations 68

4.6 Number of non-significant distances for complete sites, 5 normalizations 68

4.7 Geographic correlation for sample size 1000, 1 normalization iteration 72

4.8 Geographic correlation for complete sites, 1 normalization iteration 72

4.9 Geographic correlation for sample size 1000, 5 normalizations 73

4.10 Geographic correlation for complete sites, 5 normalizations 73

4.11 Travel correlation for sample size 1000, 1 normalization iteration 73

4.12 Travel correlation for complete sites, 1 normalization iteration 74

4.13 Travel correlation for sample size 1000, 5 normalizations 74

4.14 Travel correlation for complete sites, 5 normalizations 74

4.15 Average Inter-measure-correlation of measures 76

4.16 Size correlation for sample size 1000, 1 normalization 76

4.17 Size correlation for complete sites, 1 normalization 77

4.18 Size correlation for sample size 1000, 5 normalizations 77

4.19 Size correlation for complete sites, 5 normalizations 77

4.20 Clusters discussed 104

4.21 List of parts of speech 105

4.22 List of non-terminal labels 105

List of Figures

1.1 Abstract Distance Measure Model : $d \circ f$ 2

2.1 Swedia, Consensus Tree Map 14

2.2 Swedia, Multi-Dimensional Scaling of Trigrams measured by Jensen-Shannon divergence 15

2.3 Swedia, Composite Cluster Map 16

3.1 Dependency parse for "The dog barks." 38

3.2 Example Tree 39

3.3 Phrase-Structure Rules Extracted 39

3.4 Grandparent Phrase-Structure Rules Extracted 40

3.5 Hierarchical Cluster Dendrogram 49

3.6 Sites Before Clustering 51

3.7 Sites After A-B Merge 51

3.8 Sites After D-E Merge 52

3.9 Sites After A-B-C Merge 52

3.10 Sites After Clustering 53

3.11 Ward's method, before clustering 54

3.12 Ward's method, after A-B merge 54

3.13 Ward's method, after D-E merge 54

3.14 Ward's method, after A-B-C merge 55

3.15 Ward's method, after clustering 55

3.16 Input cluster dendrograms 55

3.17 Output consensus dendrogram 56

3.18 Spans from input trees 56

3.19 Span type frequencies (starred rows do not occur in the majority of trees) 56

3.20 Majority span types 56

3.21 Swedia, Multi-Dimensional Scaling of Trigrams measured by Jensen-Shannon divergence 59

3.22 Feature-ranking 1:1 61

3.23 Feature-ranking 1:Many 61

3.24 Feature-ranking Many:Many 62

4.1 Dendrogram With Jensen-Shannon measure and trigram features, 1 normalization, 1000 samples 80

4.2 Dendrogram With Jensen-Shannon measure and trigram features, 5 normalizations, 1000 samples 81

4.3 Dendrogram With R^2 measure and phrase-structure-rule features, 1 normalization, complete sites 82

4.4 Dendrogram with cosine measure and trigram features, 5 normalizations 83

4.5 Consensus Tree for 1000-samples and 1 normalization 85

4.6 Consensus Tree for full site comparison and 1 normalization 86

4.7 Consensus Tree for 1000-samples and 5 normalizations 87

4.8 Consensus Tree for 1000-samples and 1 normalization, Mapped 88

4.9 Consensus Tree for full site comparison and 1 normalization, Mapped 89

4.10 Consensus Tree for 1000-samples and 5 normalizations, Mapped 90

4.11 Blue Cluster 91

4.12 Red Cluster 91

4.13 Yellow Cluster 91

4.14 Cyan Cluster 92

4.15 Orange Cluster 93

4.16 Composite Cluster Map for 1000-sample, 1 normalization 95

4.17 Composite Cluster Map for complete sites, 1 normalization 96

4.18 Composite Cluster Map for complete sites, 5 normalizations 97

4.19 Jensen-Shannon measure with trigram features, 1000-sentence sampling and 1 round of normalization 99

4.20 Jensen-Shannon measure with trigram features, 1000-sentence sampling and 5 rounds of normalization 100

4.21 R^2 measure with phrase-structure-rule features, full-site comparison and 1 round of normalization 101

4.22 cluster A $\Leftrightarrow$ cluster B, trigram features 107

4.23 cluster A $\Leftrightarrow$ cluster C, trigram features 107

4.24 cluster A $\Leftrightarrow$ cluster D, trigram features 107

4.25 cluster B $\Leftrightarrow$ cluster C, trigram features 108

4.26 cluster B ⇔ cluster D, trigram features 108
4.27 cluster C ⇔ cluster D, trigram features 108
4.28 cluster A ⇔ cluster B, trigram features with overuse normalization 110
4.29 cluster A ⇔ cluster C, trigram features with overuse normalization 110
4.30 cluster A ⇔ cluster D, trigram features with overuse normalization 110
4.31 cluster B ⇔ cluster C, trigram features with overuse normalization 111
4.32 cluster B ⇔ cluster D, trigram features with overuse normalization 111
4.33 cluster C ⇔ cluster D, trigram features with overuse normalization 111
4.34 cluster A ⇔ cluster B, leaf-ancestor path features 113
4.35 cluster A ⇔ cluster C, leaf-ancestor path features 113
4.36 cluster A ⇔ cluster D, leaf-ancestor path features 113
4.37 cluster B ⇔ cluster C, leaf-ancestor path features 114
4.38 cluster B ⇔ cluster D, leaf-ancestor path features 114
4.39 cluster C ⇔ cluster D, leaf-ancestor path features 114
4.40 cluster A ⇔ cluster B, leaf-head features 115
4.41 cluster A ⇔ cluster C, leaf-head features 115
4.42 cluster A ⇔ cluster D, leaf-head features 115
4.43 cluster B ⇔ cluster C, leaf-head features 116
4.44 cluster B ⇔ cluster D, leaf-head features 116
4.45 cluster C ⇔ cluster D, leaf-head features 116
4.46 cluster A ⇔ cluster B, phrase-structure rule features 117
4.47 cluster A ⇔ cluster C, phrase-structure rule features 117

4.48 cluster A ⇔ cluster D, phrase-structure rule features 118

4.49 cluster B ⇔ cluster C, phrase-structure rule features 118

4.50 cluster B ⇔ cluster D, phrase-structure rule features 118

4.51 cluster C ⇔ cluster D, phrase-structure rule features 119

5.1 Suffix marking for partitive 124

5.2 Proper-Noun Articles 126

5.3 Indefinite Article for Proper Nouns: First Names 127

5.4 Proper-Noun Articles 128

5.5 Simultaneous possessive and determiner in noun phrase in Danish, and at one time Southwest Sweden 129

5.6 Possessive formed of Possessive Pronoun and Proper Noun 129

5.7 Proper-Noun Possessives 130

5.8 Double Indefinite 131

5.9 Double indefinite (post-adjectival articles) 132

5.10 Double definite (Sweden and Norway) 133

5.11 Single Indefinite (Denmark) 133

5.12 Single definite suffix (Iceland) 133

5.13 Single definite suffix with combined adjective (Northern Sweden) 134

5.14 Double definite (and combined adjectives) 135

5.15 Apparent Cleft 136

5.16 Apparent Cleft with adverb expressing speaker attitude 136

5.17 Factors 1 and 2 of Swedish vowels 138

5.18 Factors 3 and 4 of Swedish vowels 139

5.19 Factors 5 and 6 of Swedish vowels 140

1

Introduction

This dissertation establishes the utility and reliability of a statistical distance measure for syntactic dialectometry, expanding dialectometry's methods to include syntax as well as phonology and the lexicon. It is a continuation of my previous work (Sanders 2007, Sanders 2008) and earlier work by Nerbonne and Wiersma (2006), the first statistical measure of syntax distance. These pioneering studies explored this measure, but failed to compare it to established results in dialectology to see if the new method reproduces them. This dissertation does so, as well as investigating a number of variant measures and feature sets. It uses Swedish dialect data as a basis for investigation.

Dialectology is the study of linguistic variation (Chambers and Trudgill 1998). Its goal is to characterize the linguistic features that separate language varieties. The tools that it uses to do this include isoglosses—geographic descriptions of a particular linguistic variable–as well as traditional phonological and syntactic analyses of dialect phenomena. Traditional dialectology predates sociolinguistics, but has adopted many of its tools so that it has become in some ways a subfield of sociolinguistics.

Dialectometry is a subfield of dialectology that uses mathematically sophisticated methods to extract and combine linguistic features (Séguy 1973). Its focus is the manipulation of large data sets in a uniform way, characterizing the differences between regions in a gradient and statistically sound way. As a result, in recent years most work in the field has been computational linguistic, largely focused on phonology, starting with Kessler (1995), followed by Nerbonne and Heeringa (1997) and Nerbonne and Heeringa (2001). Heeringa (2004) provides a comprehensive review of

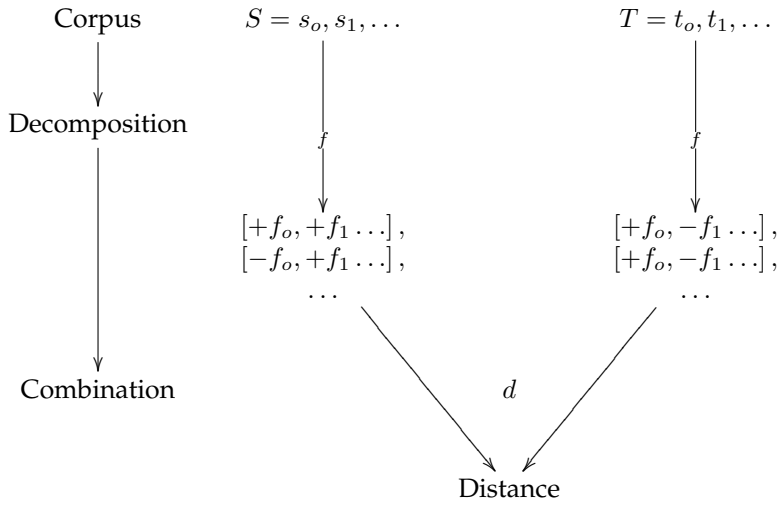


Figure 1.1: Abstract Distance Measure Model : $d \circ f$

phonological distance in dialectometry as well as some new methods.

This dissertation compares the results of the syntactic distance measure with syntactic dialectology, both in the form of the traditional Swedish dialect regions as well as analysis of syntactic dialect features. It also compares the results to phonological dialectometry's results on Swedish.

1.1 Overview of Dialectometry

In dialectometry, a distance measure can be defined in two parts: first, a method of decomposing the data into features that capture linguistic properties, and second, a method of combining the features from two corpora to produce a single number. The decomposition method can be thought of as "feature extraction" for any number of feature sets, while the combining method can be thought of as the "distance", since it represents differences as a single number. Figure 1.1 gives an overview of how the model works. The input consists of two corpora; each item in each corpus is decomposed into a set of features extracted by f. The resulting set of features are then compared by d, which combines them into a single number: the distance.

Dialectometry has focused on phonological distance measures, while syntactic measures have

remained undeveloped. The most important reason for this focus is that it is easier to define a distance measure on phonology. In phonology, it is easy to collect corpora consisting of identical word sets. Then these words decompose to segments and, if necessary, segments further decompose to phonological features. This decomposition is straightforward and based on Chomsky and Halle (1968). For combination, string alignment, or Levenshtein distance (Levenshtein 1965), is a well-understood algorithm used for measuring changes between any two sequences of characters taken from a common alphabet. Levenshtein distance is simple mathematically, and has the additional advantage that its intermediate data structures are easy to interpret as the linguistic processes of epenthesis, deletion and metathesis. Other methods have been proposed; Kondrak (2002) give several simpler alternatives to Levenshtein distance and Sanders and Chin (2006) and Hinrichs and Zastrow (2007) give two different statistical measures. However, Levenshtein distance remains dominant for currently available corpora because it maintains the best balance between required corpus size and quality of results.

These things are not possible with syntactic distance: neither matched sentences nor a single obvious function for decomposition and combinations exist. Matched sentences could in theory be collected, but the number of possible interview responses in syntax is so much larger that the required time and number of informants would be correspondingly much greater. For example, in the Survey of English Dialects (Orton and Halliday 1963), phonological items were elicited by asking the interviewee to answer a question: "cow" is the standard English answer to "What is the animal that you get milk from?". This method avoids priming the interviewee with the interviewer's pronunciation. However, It does not always have the desired effect, even for phonology: for the item "newt", the responses "newt", "ewt" and "eft" are all comparable phonologically, but a response like "salamander" is not. This problem is exponentially worse for syntax: an interview question that is sufficiently abstract to avoid priming a particular structure has a low chance of eliciting that very structure. For example, a prompt such as "I took your drink. What do you say to me?" has a low chance of eliciting sentences that exemplify the differing English orders of direct and indirect objects such as "give me it" versus "give it to me" or "give it me".

The specification of functions for decomposition and combination of syntax faces another problem. Although many decomposition and combination methods can be proposed, the standard

syntactic theories cannot be practically used. For example, the parsers in this dissertation use probabilistic phrase structure rules or dependencies to represent the grammar of a language. This is typical for parsers in computational linguistics, but it means that their output, and by extension the features based on their output, is quite different from the lexical representations of minimalism. In order to decompose sentences to minimalist features, a broad-coverage minimalist parser would be required. Since such a parser does not yet exist, it is impossible to use the minimalist syntactic structure in the same way that phonological dialectometry uses distinctive features, for example.

For similar reasons, methods from lexical dialectology are not simple to adapt to syntax. There are two problems: first, lexical feature extraction ranges from trivial to easy, so there are no applicable techniques for feature extraction. Second, there has been little work on distance measure specifically for lexical dialectometry. Lexical distance typically uses a non-specific method such as Goebl's GIW, described in chapter 3; see for example Spruit et al. (2006).

Syntax and Dialectometry

Because of the preceding two reasons, syntax is a relatively undeveloped area in dialectometry. Currently, the literature lacks a generally accepted syntax measure. Unfortunately, approaching the problem by copying phonology is not a good solution; there are real differences between syntax and phonology that mean phonological approaches do not apply. For example, there are fewer differences to be found in syntax, and they occur more sparsely. Because dialectology has traditionally worked with fairly small corpora, and because of the difficulty of collecting syntactic data, most surveys cover even fewer syntactic variables than phonological ones. There are two approaches to remedy this. The first manually enhances the differences that do exist in small, carefully collected corpora; the second switches to larger, non-survey corpora and uses statistical methods to find differences.

The first approach is proposed by Spruit (2008) for analyzing the Syntactic Atlas of the Dutch Dialects (Barbiers et al. 2005), is to continue using small dialectology corpora and manually extract features so that only the most salient features are used. Then a sophisticated method of combination such as Goebl's Weighted Identity Value (WIV), described in chapter 3, and by Goebl (2006),

can be used to produce a distance. WIV is more complex mathematically than Levenshtein distance, and operates on any type of linguistic feature. However, manual feature extraction requires that the dialect situation be understood first. In other words, traditional dialectology methods must be used to find interesting features before dialectometry can proceed. This negates the usual advantages of dialectometric methods in providing rapid analysis in knowledge-poor environments. Manual feature extraction is also subject to bias from the dialectologist: the best-known features are most likely to become the best manual features, passing over the rarely occurring and previously unknown features that might actually be the best indicators of a particular dialect.

This first approach ignores the specific properties of the syntax distance problem. Given a large corpus, manually defined features will have less coverage than the automatically extracted features used by the second, statistical approach. Furthermore, automatically extracted features are easy to define for syntax. This dissertation covers part-of-speech trigrams, leaf-ancestor paths, and leaf-head paths over nodes, but many variations on these features are possible, such as lexical trigrams, lexicalized leaf-ancestor paths, or arc-head paths. Methods from other syntactic work in computational linguistics could apply too: supertags (Joshi and Srinivas 1994), convolution kernels (Collins and Duffy 2001) or any number of simpler features such as tree height, number of nodes, or number of words.

The problem for the statistical approach is not defining a feature set. The problem is defining a good feature set. This is the reason that the statistical approach uses large corpora: with enough data, statistically significant comparisons can be made between the different features; the highest ranked ones can be discovered automatically rather than manually. Fortunately, the typical syntactic corpus is larger than a phonological corpus because the annotation work is easier; much of the syntactic annotation can be generated automatically.

Even with a feature set defined, a distance measure still requires a method of combining features to find a distance. One such method, a simple statistical measure called R, has been proposed by Nerbonne and Wiersma (2006) based on work by Kessler (2001). At present, however, R has not been adequately shown to detect dialect differences. A small body of work suggests that it does, but as yet there has not been a satisfying correlation of its results with existing results from the dialectology literature on syntax.

Nerbonne & Wiersma's first paper used part-of-speech trigram features as a proxy for syntactic information and R for syntax distance together with a test for statistical significance(Nerbonne and Wiersma 2006). Their experiment compared two generations of Norwegian L2 speakers of English. They found that the two generations were significantly different, although they had to normalize the trigram counts to account for differences in sentence length and complexity. However, showing that two generations of speakers are significantly different with respect to R does not necessarily imply that the same will be true for other types of language varieties. Specifically, for this dissertation, the success of R on generational differences does not imply success on dialect differences.

I addressed this problem (Sanders 2008) by measuring R between the nine Government Office Regions of England, using the International Corpus of English Great Britain (Nelson et al. 2002); see the discussion in section 3.2. Speakers were classified by birthplace. I also introduced Sampson's leaf-ancestor paths as a feature set (Sampson 2000). I found statistically significant differences between most regions, using both trigrams and leaf-ancestor paths as features. However, R's distances were not significantly correlated with Levenshtein distances. Nor did I show any qualitative similarities between known syntactic dialect features and the high-ranked features used by R in producing its distance. As a result, it is not clear whether the significant R distances correlate either with dialectometric phonological distance or with known features found by dialectologists.

1.2 Overview of the Dissertation

The problem outlined in the previous section is that dialectometry lacks a statistical method designed for syntax which does not require the linguist to specify ad-hoc features manually. This dissertation addresses the lack directly by applying the method to a dialect corpus, then comparing the results to existing syntactic dialectology literature of Swedish, as well as phonological work using established dialectometry methods. In addition, it tests variations of the experimental parameters in order to identify the highest-performing parameters. In summary, this analysis allows future dialectometry studies to include syntactic as well as phonological analyses, having an idea of the best method and parameters to use.

There are three research questions that must be answered to determine the reliability of this measure. They are given in chapter 2. First, does the measure agree with the results of dialectology? Previous work has not addressed this question, but it is crucial that a new measure reproduce the results from previous linguistic work. To answer this question, the Swedish dialect distance results will be processed in a number of ways so that they are comparable to previous dialect work on Swedish in multiple ways.

Second, which parameter variations produce the best agreement with dialectology work? Both the distance measure and feature set can be varied, as well as a number of other parameter settings, mostly dealing with controlling for the effects of corpus size. The distance measures include simple measures like R, which is a sum of differences, more complex variants such as Jensen-Shannon divergence, which is a sum of logarithmic differences, and cosine similarity, which models each corpus as a vector in high-dimensional space and finds the angle between two corpus vectors. Feature sets can be even more varied, although all the feature sets discussed here assume that the word is the basic unit of syntactic analysis and that words are naturally grouped into sentences. Some example feature sets are part-of-speech trigrams, which are simply triples of parts of speech. Leaf-ancestor paths and leaf-head paths use the syntactic structure of the sentence, with leaf-ancestor paths based on constituent grammars (phrase-structure grammars) and leaf-head paths based on dependency grammars.

Third, does the measure agree with the results of phonological dialectometry? Agreement is not required; phonological and syntactic dialect boundaries may disagree, but they are more likely to agree than disagree, so if the two dialectometric measurements agree, then this inspires confidence on the new method based on the old method's reliability.

To answer the three research questions, I start with the statistical method described in the previous section with the parameter variations described above in chapter 3. To make sure that the results are comparable to previous dialectology, I use the dialect corpus Swediasyn, which is a transcription of interviews recorded in villages throughout Sweden. The interviewees were balanced between older and younger men and women. To generate features from the Swediasyn, a good deal of processing is required; the corpus is a transcription with no syntactic annotation. To annotate the Swediasyn, I use a number of automatic annotators, trained on Talbanken, a corpus of spoken and

written Swedish. However, Talbanken does not include dialect sources, so error is expected during the annotation process. After annotation, feature generation is straightforward: transformation of parse trees and other annotations. Because automatic annotators should make identical mistakes when annotating identical dialect structures, the resulting features should contribute usefully to distance, despite being incorrect linguistically.

After measuring distances between the interview sites, a number of analytic methods are applied to the distances so that they can be compared to dialectology work. The methods are a test of significance, a test of correlation, cluster dendrograms and consensus trees, composite cluster maps, multi-dimensional scaling, and feature ranking. The tests of significance and correlation represent the distances' trustworthiness and ability to match dialectology's assumptions, respectively. The consensus trees, and multi-dimensional scaling both produce maps. These maps allow the linguist to visually compare the results with traditional region maps. In the same way, composite cluster maps allow visual comparison of the results to isogloss bundles from dialectology. Finally, feature ranking allows the linguist to view the features that contribute most to separating two regions. These features can be compared to the dialect phenomena cataloged by dialectologists.

The results in chapter 4 are presented in the same order as their corresponding analysis appear in 3. The dissertation concludes with discussion in chapter 5. Here, I compare the results to the dialectology and phonological dialectometry of Swedish. Then I discuss the relation of this work to previous work in syntactic dialectology, detailing its contribution to the field. I finish by presenting avenues for future work: with a statistical measure of dialect distance, dialectometry can analyze syntactic features as well as phonological and lexical ones, producing more complete analyses.

2

Questions

The state of syntax measures in dialectometry described above leaves several research questions unresolved. It is not yet clear whether R is a good measure of syntax distance. Previous results have shown that it can obtain significant distances, but has either failed to do so reliably, as in my work on British English (Sanders 2008), or has not compared traditional dialect areas, as in Nerbonne and Wiersma (2006). Neither study showed that a statistical method could adequately reproduce existing knowledge about some dialect area, which is necessary before R, and statistical methods as a whole, can contribute to dialectometry's study of syntax.

This leads to the first question: will the features found by dialectologists agree with the highly ranked features used by a statistical method for classification? I will investigate this question by comparing statistical dialectometry results to the syntactic dialectology literature on Swedish. A secondary, but related question is whether the regions of Sweden accepted by dialectology will be reproduced by a statistical method. For example, my previous research on British English reproduced the well-known North England-South England dialect regions. However, this dissertation eliminates the corpus variability in that research, where a forty-year gap separated the phonology and syntax corpora, and the syntax corpus was not collected with dialectology in mind (Sanders 2008). With a corpus collected for the purpose of dialect research, and with a phonological corpus transcribed from the same interviews, more precise comparisons should be possible, both between regions and between syntax and phonology.

A secondary question, relevant once the utility of a statistical measure for syntax is established,

is what variations of the two functions comprising the measure produce the best results. This involves variation of both the feature extraction function and the distance function. Choice of feature set is almost as important as choice of distance. My previous work on British English showed that leaf-ancestor paths provide a small advantage over part-of-speech (POS) trigrams, presumably by capturing syntactic structure higher in the parse tree. And, whereas development of a statistical distance measure is difficult, new feature sets can be developed relatively quickly. In this dissertation, I evaluate several feature sets besides POS trigrams and leaf-ancestor paths, such as phrase structure rules, leaf-head paths, and lexical trigrams. I also evaluate variants of these feature sets, for example varying the POS tagger or POS tag set. I also evaluate combined feature sets.

Feature sets can be evaluated by comparing performance of different feature sets on a fixed corpus and with a fixed distance measure. Here, performance is measured using the same criteria as for distance measures: the number of significant distances between interview sites and the similarity of the results to those found by dialectologists.

Besides feature sets, this dissertation evaluates a number of measures beyond the R of previous work, such as Kullbeck-Leibler divergence and cosine dissimilarity. R is one way to aggregate features that are created by decomposing sentences. It treats features as atomic, and does not manipulate them in any syntax-specific ways. As such, R differs from Goebl's WIV only in being designed for larger feature sets and larger corpora. Both assume that independent, atomic features derived from a sentence can adequately capture dialect differences. If this is not the case, then a more syntax-aware way of comparing individual features will be needed.

A final question is whether the syntactic dialectometry practiced here agrees with phonological dialectometry on the same corpus. Unlike the previous questions, which use agreement between syntactic dialectometry and dialectology, there is no *a priori* reason to expect syntax/phonology agreement; it is quite possible that phonological features create one set of boundaries while syntactic feature create another set. However, agreement between the two would be further evidence for that statistical methods are useful for syntactic dialectometry.

2.1 Question 1 : Agreement with Dialectology

The first question is whether a statistical dialectometry measure agrees with dialectology. On closer inspection, this question covers a number of more specific questions, each dealing with a specific comparison to dialectology. First, and most important, is whether the features that it counts most important are the same as the features discussed in the dialectology literature. Three other questions are whether regions, region boundaries, and distances found by this measure agree with dialectology. Therefore, question 1 has a four-part answer: agreement between dialectometry and dialectology on regions, boundaries, distances, and features.

First, however, these terms from dialectology must be defined precisely. Then the methods used to compare the dialectometry results with dialectology can be developed.

Definition of Dialectology Terms

Definition of terms from dialectology is appropriate here, along with an explanation of how they fit together. The basic unit in dialectology is the feature, such as "pronunciation of the word 'cow' " or "adjective placement in noun phrases". During analysis, the linguist may suspect that a certain variant of a feature is characteristic of a particular region, but more information, usually from a survey, is needed to make certain.

Given a survey or other source of geographical mapping information, a boundary for a feature can be drawn. This boundary is called an isogloss. For simple cases, isoglosses are usually simple to determine, giving a clear line between two dialects. On the other hand, complicated cases lead to more complicated geometry; for example, a few occurrences of a feature variant can be stranded in the middle of the other variant.

If a number of isoglosses coincide, they form an isogloss bundle, which separates one region from another. Isogloss bundles are simple in theory, but in practice they are difficult to find because isoglosses rarely coincide perfectly. In practice, undisputed isogloss bundles only occur between well-known dialects, such as the boundaries between Low and High German or Northern and

Southern English of England. In cases where more precision is required, there is not usually a sufficient number of coincident isoglosses. Even though there may be plenty of isoglosses in the area, isoglosses so rarely coincide that only a few may be construed as forming an isogloss boundary.

Dialectology does not have a clear equivalent to dialectometry's distance. The closest analog is size of isogloss bundle; dialect maps typically indicate size of isogloss bundle by thickness of boundary line. Additionally, regions that have many specific features known in the dialectology literature can be inferred to be distant from the rest.

Features

The first aspect of dialectology to compare is the feature. To match the features of dialectology to the features that a statistical dialectometric method uses to produce a distance, I first need to find discussion of Swedish dialect features in the dialectology literature. For example, Rosenkvist (2007) discusses the South Swedish apparent cleft. Here, the sentence contains an embedded clause with similar surface appearance to a true cleft. Unlike a true cleft, however, there is no clefted constituent in the matrix clause. The apparent cleft appears in southern Sweden, but its precise distribution is not known; Rosenkvist finds some uses everywhere except Norrland (northern Sweden), but finds heaviest use in the former Danish provinces in the south.

Next the feature should be expressed formally. This formal description can then be translated to the format representation used by the dialectometry. Again, these would be the same ideally, but the dialectology study may not be complete; for example, Rosenkvist's 2007 paper does not yet include a syntactic analysis. And even in cases where the dialectology gives a formal description, the syntactic features of dialectometry, for example those described in the next chapter, are based on more primitive formalisms at present. Therefore the translation may lose information. Once translated, the features discussed by dialectologists should appear in the high-ranked features on which the statistical dialectometry method bases its distance.

In the apparent cleft example, the apparent cleft is realized as an additional use of the word *som*, ordinarily a complementizer. Typically, the next step is to identify the minimalist structure for this, but Rosenkvist's 2007 paper does not yet provide this analysis. Although there is no structure

to translate to a phrase-structure skeleton, his analysis provides enough clues to produce some features directly. Part-of-speech n-grams are easiest; he mentions that his corpus search used the strings *det är som* ("It is that") and *det är bara som* "It's just that". These words only need part-of-speech annotation to be n-gram features. Leaf-head paths can also use these parts of speech for the local dependencies between *det*,*är*, and *som*. Rosenkvist also mentions some syntactic properties of apparent clefts that are useful for specifying leaf-head path features: the subject of the *som*-clause must be a pronoun, so we should expect to see leaf-head paths of the form *ROOT-som-PRON* in the regions that have the apparent cleft.

Once dialectometric features have been specified from some linguistic analysis, the analysis consists of the following questions: in what regions do these features appear? Do these regions match the expected distribution (if any) from the linguistic analysis? How much do the features contribute to distance from other regions? If there are other features that contribute more, what are they?

Isogloss Boundaries

Isogloss boundaries are intermediate in complexity between features unspecified for location and regions demarcated by isogloss bundles. For the purposes of this dissertation, however, there is not much difference between a feature with some documented locations and an isogloss boundary. An isogloss makes the regions of interest clearer, but it is a difference in degree and not in quality. The real difference in analysis occurs when dialectology has identified an isogloss bundle.

Isogloss Bundles

Isogloss bundles compare straightforwardly to dialectometry, once regions have been identified from the dialectometric distances between sites. There are two primary methods: hierarchical clustering and multi-dimensional scaling. Neither method is perfect; as with isogloss bundles, some human input is still needed to determine whether an inter-region boundary truly exists at some point.

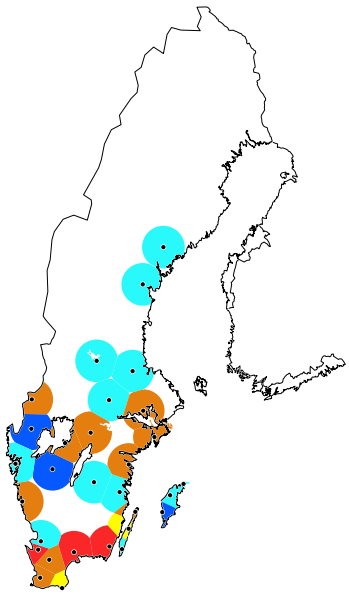


Figure 2.1: Swedia, Consensus Tree Map

Hierarchical clustering produces well-delineated regions by recursively merging sites into regions, at the cost of some uncertainty—the results tend to vary quite a bit from feature set to feature set. Only clusters that persist between results from multiple feature sets should be considered valid. Consensus trees aggregate multiple cluster dendrograms into a stable tree; see figure 2.1 for an example. However, because of the recursive, nested nature of the grouping, there can still be a question of which level of nesting is appropriate to treat as a region.

In contrast, multi-dimensional scaling (MDS) is a mathematical transformation of the high-dimensional space created by measuring distances between all sites in the corpus; see figure 2.2 for an example and section 3.4 for a complete discussion. Although MDS does not produce spurious information, its results are often hard to analyze because it produces boundaries of varying strength. Very different regions stand out, but similar regions appear similar even if they contain some differences. This similarity can make it difficult to decide whether an area should be considered one region or two.

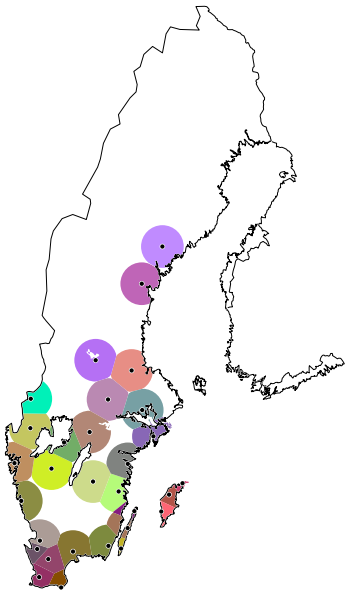


Figure 2.2: Swedia, Multi-Dimensional Scaling of Trigrams measured by Jensen-Shannon divergence

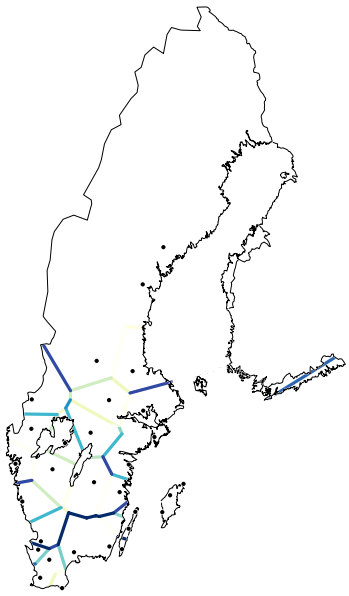


Figure 2.3: Swedia, Composite Cluster Map

Once both dialectologic and dialectometric regions have been identified, comparison is straightforward. Each region can be checked for overlap—regions with a greater overlap area are better matches.

Distances

Although comparing distances from dialectometry to qualitative research in dialectology is possible, it is not very precise, because the dialectometric distances must first be translated to something like the isogloss bundles of dialectometry. Composite cluster maps provide this translation by drawing dark boundaries when large distances separate regions; see figure 2.3 and discussion in section 3.4. Alternatively, statements like "in general, Southern Swedish is syntactically identical to Standard Swedish" (Rosenkvist 2007) can be construed as saying, roughly, that there is very little distance between Southern and Standard Swedish. Ultimately, though, the distances from a quantitative analysis do not have a clear analog in qualitative analyses.

2.2 Question 2 : Variations on the Measure

The second question of this dissertation reflects the fact that the distance measures in dialectometry have two parts. The first part is the function used to extract features from a corpus and the second is the distance measure that produces a distance between the features of two corpora. This dissertation investigates a number of implementations for both functions. The question is which combination provides the best performance, as measured by agreement with dialectology.

Specification of feature sets is not difficult; feature sets are easier to create than distance measure algorithms, as discussion of distance measures below will show. In addition, feature sets are easier to combine and to tweak. The real problem is not in specification of feature sets, but that new feature sets must be evaluated, since it is not currently possible to produce features based on a linguistic theory as with phonology's distinctive features.

For example, in previous work, I showed that leaf-ancestor paths have a small advantage of trigrams (Sanders 2007) in terms of finding significant distances. Therefore, Question 2 breaks into two smaller questions: (1) how can new variations be proposed? and (2) how can they be evaluated? However, before these questions are explored, an definition of terms related to distance measures is in order.

Definition of dialectometry terms

There are several terms related to distance in mathematics. In order from least restrictive to most restrictive, they are 'divergence', 'dissimilarity' and 'distance'. In this dissertation, a 'measure' is used to refer to any of these three functions. All three kinds of functions must always return positive numbers, and only return 0 for corpora that are equal. A symmetric function returns the same number whether measuring from point X to point Y or from point Y to point X. The triangle inequality means that distance from point X to point Y plus point Y to point Z is at least as long as traveling straight from point X to point Z. In other words, it means that it will always be longer to take the two-leg path than to take the single-leg path. Equations 2.1-2.4 list the properties formally.

$$d(x, y) \geq 0 \tag{2.1}$$

$$d(x, y) = 0 \text{ iff } x = y \tag{2.2}$$

$$d(x, y) = d(y, x) \tag{2.3}$$

$$d(x, y) + d(y, z) \geq d(x, z) \tag{2.4}$$

A divergence satisfies equations 2.1 and 2.2: it is always positive and only zero when two sites are equal. It is less restrictive than the other two kinds of measures, and is the only one that can capture the common dialect situation where speakers of dialect X can understand speakers of dialect Y better than speakers of Y understand those of X. Unfortunately, the methods used for aggregate comparison in dialectology, such as hierarchical dendrograms and multi-dimensional scaling (MDS), require more restrictive measures. Specifically, dissimilarities are divergences that are in addition symmetric. Dissimilarities can be used in hierarchical dendrograms and MDS because multiple dissimilarity comparisons can be mapped into distance space by situating each pair of sites in its own orthogonal dimension. This high-dimensionality avoids violating the triangle inequality. Finally, distances are dissimilarities that additionally satisfy the triangle inequality without special consideration, so multiple pairwise comparisons can inhabit the same dimensions. However, this is not necessary for the analyses in this dissertation.

Therefore, the measures described in the rest of the dissertation will be dissimilarities, but not necessarily distances. In the rest of the dissertation, 'distance' will usually be used as a generic term to refer to a dissimilarity; exceptions where the term 'distance' implies all three properties will be noted. In addition, some of the dissimilarities have common names that contain other terms. For example, Kullback-Leibler divergence is augmented here to behave as a dissimilarity, but it retains its original name when mentioned.

Feature Sets

New feature sets are easy to propose. All that is needed is some way to condense or divide the information about the sentence into symbols that can be used as input to a statistical distance measure. Specifically, the feature sets used in this dissertation use per-word information, word-order information, and syntactic information. They attach some information from the constituent tree or dependency graph to each word, dividing the information according to the word's position in the sentence. Trigrams attach the leaves to each word, along with the leaves to the left and right. Leaf-ancestor paths attach vertical slices of the tree to each word. Leaf-head paths attach the path to the root to each word.

Feature sets that use other information might also be useful; convolution kernels give a single number that captures the difference between two trees (Collins and Duffy 2001); a similar feature that captures aspects of a single tree such as depth, branching degree or homogeneity might be useful. Besides this, there are numerous simple features used in other computational linguistic work that attempt to capture the most important characteristics of a sentence in a simple, ad-hoc way, such as the first or last n words of a sentence, a certain number of words surrounding the predicate, or sentence length.

Even before looking at results, it seems that each of these has its own advantages and disadvantages. Leaf-ancestor paths capture upper structure of the constituent parse, but no left and right context. Leaf-head paths capture some of the sentence structure, but some of the surrounding context as well. Trigrams capture only immediate left-right context, but include word order information. They are also less influenced by annotator error since they require only part-of-speech annotation.

Because evaluation of feature set performance is necessarily evaluation of the overall combination of feature set and measure, the previously discussed measures of agreement with dialectology should all be used as measures of performance. With the distance measure held constant, the different feature sets can be evaluated against one another.

Before comparison, though, the distances produced for a given combination of feature set and measure must be checked for significance. For example, a very sensitive combination could be

inappropriate for small data sets if it can only achieve significance with large data sets. The significance test ensures that subsequent evaluation is valid.

Distance Measures

Of the measures considered in this dissertation, R and R^2 have been tested in previous work. R is quite simple; it is a sum of differences of features. It treats features as opaque symbols; it is not necessarily limited syntax. Perhaps because of its simplicity, R performs more consistently than other measures tested in this dissertation: It gives significant results across a larger variety of feature sets than more complicated measures do.

There are two obvious directions to explore when creating a distance measure to replace R. The first direction is to address R's simplicity by defining a more complex measure that uses sophisticated ways to measure difference over still-opaque symbolic features. The second direction is to address R's ignorance of syntax by defining a measure with specific knowledge of syntax. Finding candidates for the first direction is easier, given the number of statistical measures commonly used in computational linguistics. Additionally, the dialectometric model that divides a measure into distance measure and feature set is powerful enough that most syntax-specific knowledge can be represented in terms of features instead of integrated into the distance measure's algorithm.

Indeed, this makes syntax-aware measures difficult to specify—they must incorporate knowledge of syntax in a way that cannot be reified as features. Unlike dialect surveys of phonology, dialect interviews do not consist of aligned lists of sentences. That means that pairwise sentence-to-sentence comparison are impossible; comparison must occur at a lower level. This constraint makes it difficult to encode any useful awareness of syntax into a syntax-aware distance measure that cannot be easily represented in the feature set for a syntax-ignorant measure instead.

It is so difficult to define a useful syntax-aware distance measure that none are presented in this dissertation. Syntax awareness is restricted to the feature sets. However, a number of more complicated statistical measures similar to R are presented. Evaluation of the distance measures proceeds similarly to evaluation of feature sets; results for various measures are compared, holding the feature set constant. The results are checked for significance, then for agreement with dialectology.

2.3 Question 3 : Agreement with Phonological Dialectometry

Finally, agreement with phonological dialectometry is a useful indicator of quality. Agreement with phonology indicates a good feature set, but cannot indicate a bad feature set. Phonological boundaries need not agree with syntactic boundaries, but it seems *a priori* likely that they do. Note that agreement with phonology has the reverse implication of statistical significance—a test for significance can only indicate a bad feature set, not prove a good feature set.

There is very little phonological dialectometry for Swedish, so this comparison may not be valid yet. The only published paper, to my knowledge, is Leinonen (2008). Leinonen has extended this work to a dissertation, which is currently unpublished.

3

Methods

This chapter contains four sections. The first discusses related work, dialectometry since its development in the middle of the 20th century, starting with Séguy (1973) and continuing with Goebl (2006). The second section discusses the previous work on statistical methods for syntactic dialectometry, as well as the feature sets and distance measures developed in this dissertation. The third and fourth sections cover the application of the work to Swedish dialects. Specifically, the third section deals with input analysis: which Swedish corpora were used and which annotators served as a basis for extracting features. The fourth section covers output analysis, detailing the methods used to process the distances between interview sites so that they can be compared to the dialectology literature.

3.1 Related Work

Séguy

Measurement of linguistic similarity has always been a part of linguistics. However, until Séguy (1973) dubbed a new set of approaches 'dialectometry', these methods lagged behind the rest of linguistics in formality. Séguy's quantitative analysis of Gascogne French, while not aided by computer, was the predecessor of more powerful statistical methods that essentially required the use of computer as well as establishing the field's general dependence on well-crafted dialect surveys that

divide incoming data along traditional linguistic boundaries: phonology, morphology, syntax, etc. This makes both collection and analysis easier, although it requires more work to combine separate analyses to produce a complete picture of dialect variation.

The project to build the Atlas Linguistique et Ethnographique de la Gascogne, which Séguy directed, collected data in a dialect survey of Gascogne which asked speakers questions informed by different areas of linguistics. For example, the pronunciation of 'dog' (*chien*) was collected to measure phonological variation. It had two common variants and many other rare ones: [kãn], [kã], as well as [ka], [ko], [kano], among others. These variants were, for the most part, known by linguists ahead of time, but their exact geographical distribution was not.

The atlases, as eventually published, contained not only annotated maps, but some analyses as well. These analyses were what Séguy named dialectometry. Dialectometry differs from previous attempts to find dialect boundaries in the way it combines information from the dialect survey. Previously, dialectologists found isogloss boundaries for individual items. A dialect boundary was generated when enough individual isogloss boundaries coincided. However, for any real corpus, there is so much individual variation that only major dialect boundaries can be captured this way.

Séguy reversed the process. He first combined survey data to get a numeric score between each site. Then he posited dialect boundaries where large distances resulted between sites. The difference is important, because a single numeric score is easier to analyze than hundreds of individual boundaries. Much more subtle dialect boundaries are visible this way; where before one saw only a jumble of conflicting boundary lines, now one sees smaller, but consistent, numerical differences separating regions. Dialectometry enables classification of gradient dialect boundaries, since now one can distinguish weak and strong boundaries. Previously, weak boundaries were too uncertain.

However, Séguy's method of combination is simple both linguistically and mathematically. When comparing two sites, any difference in a response is counted as 1. Only identical responses count as a distance of 0. Words are not analyzed phonologically, nor are responses weighted by their relative amount of variation. Finally, only geographically adjacent sites are compared. This is a reasonable restriction, but later studies were able to lift it because of the availability of greater computational power. Work following Séguy's improves on both aspects. In particular, Hans Goebl

developed dialectometry models that are more mathematically sophisticated, while retaining the survey-style small feature set.

Goebl

Hans Goebl emerged as a leader in the field of dialectometry, formalizing the aims and methods of dialectometry. His primary contribution was development of various methods to combine individual distances into global distances and global distances into global clusters. These methods were more sophisticated mathematically than previous dialectometry and operated on any features extracted from the data. His analyses have used primarily the Atlas Linguistique de Français.

Goebl (2006) provides a summary of his work. Most relevant are the measures Relative Identity Value and Weighted Identity Value. They are general methods that are the basis for nearly all subsequent fine-grained dialectometrical analyses. They have three important properties. First, they are independent of the source data. They can operate over any linguistic data for which they are given a feature set, such as the one proposed by Geršić (1971) for phonology. Second, they can compare data even for items that do not have identical feature sets, unlike Geršić's measure d, for example, which cannot compare consonants and vowels. Third, they can compare data sets that are missing some entries. This improves on Séguy's analysis by providing a principled way to handle missing survey responses.

Relative Identity Value, when comparing any two items, counts the number of features which share the same value and then discounts (lowers) the importance of the result by the number of unshared features. The result is a single percentage that indicates relative similarity. This percentage, when measured between all pairs of sites in a corpus, can be scaled to produce a dissimilarity. Note that the presentation below splits Goebl's original equations into more manageable pieces; the high-level equation for Relative Identity Value is:

$$\frac{\text{identical}_{jk}}{\text{identical}_{jk} - \text{unidentical}_{jk}} \tag{3.1}$$

For some items being compared j and k. In this case *identical* is

$$\text{identical}_{jk} = |f \in \tilde{\text{N}}_{jk} : f_j = f_k| \tag{3.2}$$

where $\tilde{\text{N}}_{jk}$ is the set of features shared by j and k. In other words, of the total universe of features N, both j and k must contain the feature for it to be included in $\tilde{\text{N}}_{jk}$. So if a feature occurs only in j but not in k, it will be included in N, but not in $\tilde{\text{N}}_{jk}$. This ensures that the comparison $f_j = f_k$ is always valid, where f_j and f_k are the value of some feature f for j and k respectively. *unidentical* is defined similarly, except that it counts all features N, not just the shared features $\tilde{\text{N}}_{jk}$. Here, features that occur in only j or only k contribute toward *unidentical*'s total.

$$\text{unidentical}_{jk} = |f \in \text{N} : f_j \neq f_k| \tag{3.3}$$

Weighted Identity Value (WIV) is a refinement of Relative Identity Value. This measure defines some differences as more important than others. In particular, feature values that only occur in a few items give more information than feature values that appear in a large number of items. Wiersma's (2009) normalization, covered at the end of this chapter, reuses this idea for feature ranking.

The reasoning behind this idea is fairly simple. Goebl is interested in feature values that occur in only a few items. If a feature has some value that is shared by all of the items, then all items belong to the same group. This feature value provides *no* useful information for distinguishing the items. The situation improves if all but one item share the same value for a feature; at least there are now two groups, although the larger group is still not very informative. The most information is available if each item being studied has a different value for a feature; the items fall trivially into singleton groups, one per item.

Equation 3.4 implements this idea by discounting the *identical* count from equation 3.1 by the amount of information that feature value conveys. The amount of information, as discussed above, is based on the number of items that share a particular value for a feature. If all items share the same value for some feature, then *identical* will be discounted all the way to zero–the feature conveys no

useful information. Weighted Identical Value's equation for *identical* is therefore

$$\text{identical} = \sum_f \begin{cases} 0 & \text{if} f_j \neq f_k \\ 1 - \frac{\text{agree}_{f_j}}{(Ni)w} & \text{if} f_j = f_k \end{cases} \tag{3.4}$$

The complete definition of Weighted Identity Value is

$$\sum_i \frac{\sum_f \begin{cases} 0 & \text{if} f_j \neq f_k \\ 1 - \frac{\text{agree}_{f_j}}{(Ni)w} & \text{if} f_j = f_k \end{cases}}{\sum_f \begin{cases} 0 & \text{if} f_j \neq f_k \\ 1 - \frac{\text{agree}_{f_j}}{(Ni)w} & \text{if} f_j = f_k \end{cases} - |f \in \mathrm{N} : f_j \neq f_k|} \tag{3.5}$$

where agree_{f_j} is the number of items that agree with item j on feature f and Ni is the total number of items (w is the weight, discussed below). Because of the piecewise definition of *identical*, this number is always at least 1 because f_k agrees already with f_j. This equation takes the count of shared features and weights them by the size of the sharing group. The features that are shared with a large number of other items get a larger fraction of the normal count subtracted. WIV is similar to entropy from information theory, which forms the basis of the Kullback-Leibler and Jensen-Shannon divergences described later in this chapter (Lin 1991). The difference is that WIV subtracts values from 1 to make common features less important, while entropy takes the logarithm. The result is similar, but the two divergences are theoretically more principled in directly referring to information theory.

For example, let j and k be sets of productions for the underlying English segment /s/. The allophones of /s/ vary mostly on the feature *voice*. Seeing an unvoiced [s] for /s/ is less "surprising" than seeing a voiced [z], so the discounting process should reflect this. For example, assume that an English corpus contains 2000 underlying /s/ segments. If 500 of them are realized as [z], the discounting for *voice* will be as follows:

$$\begin{aligned} identical_{/s/\rightarrow[z]} &= 1 - 500/2000 = 1 - 0.25 = 0.75 \\ identical_{/s/\rightarrow[s]} &= 1 - 1500/2000 = 1 - 0.75 = 0.25 \end{aligned} \tag{3.6}$$

Each time /s/ surfaces as [s], it only receives 1/4 of a point toward the agreement score when it matches another [s]. When /s/ surfaces as [z], it receives three times as much for matching another [z]: 3/4 points towards the agreement score. If the alternation is even more weighted toward faithfulness, the ratio changes even more; if /s/ surfaces as [z] only 1/10 of the time, then [z] receives 9 times more value for matching than [s] does.

The final value, w, which is what gives the name "weighted identity value" to this measure, provides a way to control how much is discounted. A high w will subtract more from uninteresting groups, so that *voice* might be worth less than *place* for /t/ because /t/'s allophones vary more over *place*. In equation 3.6, w is left at 1 to facilitate the presentation, but typically it is used like an ad-hoc equivalent of information gain: the linguist can give more weight to features that are believed to be salient.

3.2 Dialectometry

It is at this point that the two types of analysis, phonological and syntactic, diverge. Although Goebl's techniques are general enough to operate over any set of features that can be extracted, better results can be obtained by specializing the general measures above to take advantage of properties of the input. Specifically, the application of computational linguistics to dialectometry beginning in the 1990s introduced methods from other fields. These methods, while generally giving more accurate results quickly, are tied to the type of data on which they operate.

Currently, the dominant phonological distance measure is Levenshtein distance. This distance is essentially the count of differing segments, although various refinements have been tried, such as inclusion of distinctive features or phonetic correlates. Heeringa (2004) gives an excellent analysis of the applications and variations of Levenshtein distance. He investigated varying levels of

detail and differing feature sets. Interestingly, although he extracted features from phonetic correlates, phonological (distinctive) features, segments, and orthographic characters, the more complex features failed to give any significant improvement over simple segments. In addition, while Levenshtein distance provides much information as a classifier, it is limited because it must have a word-aligned corpus for comparison. A number of statistical methods have been proposed that remove this requirement such as Hinrichs and Zastrow (2007) and Sanders and Chin (2009), but none have been as successful on existing dialect resources, which are small and are already word-aligned. New resources are not easy to develop because the statistical methods still rely on a phonetic transcription process.

Syntactic Distance

Recently, computational dialectometry has expanded to analysis of syntax as well. The first work in this area was Nerbonne and Wiersma's (2006) analysis of Finnish L2 learners of English, followed by Sanders's (2007) analysis of British dialect areas. As explained in chapters 1 and 2, syntax distance must be approached quite differently than phonological distance. Syntactic corpora can be built quickly by automatically annotating raw text, so it is easier to build a large syntactic corpus than a phonological one; phonological annotation does not yet have a method for automatic annotation. However, automatic annotators, while faster, cannot compete with human annotators in quality of annotation. This trade-off between annotation methods leads to the principal difference between present phonological and syntactic corpora: phonology data is word-aligned, keeping varying segments relatively close, while syntax data is not sentence-aligned, meaning that variation is distributed throughout the corpus. This difference leads syntactic approaches naturally to statistical measures over large amounts of data rather than more sensitive measures that operate on small corpora.

Nerbonne and Wiersma (2006) were the first to use the syntactic distance measure described below. They analyzed a corpus of Finnish L2 speakers of English, divided by age. The first age group consisted of speakers who learned English after childhood and the second of speakers who learned English as children. Nerbonne & Wiersma found a significant difference between the two age

groups. The features that were unexpected in English contributed most to the difference; these were associated primarily with the older age group. For example, some important features for the older age group involved determiners, which English has but Finnish does not. The features showed underuse of determiners, as well as overuse, probably due to hypercorrection. Interestingly, some of these features occur in the younger age group, but not as often. Nerbonne & Wiersma analyzed this pattern as interference from Finnish; the younger age group learned English more completely with less interference from Finnish.

My subsequent work in (Sanders 2007) and (Sanders 2008) expanded on the Finnish experiment in two ways. First, it introduced leaf-ancestor paths as an alternative feature type. Second, it tested the distance method on a larger set of sites: the Government Office Regions of England, as well as Scotland and Wales, for a total of 11 sites. Each was smaller than the Finnish L2 age groups, so the permutation test parameters had to be adjusted for some feature combinations.

The distances between regions were clustered using hierarchical agglomerative clustering, as described in section 3.4. The resulting tree showed a North/South distinction with some unexpected differences from previously hypothesized dialect boundaries; for example, the Northwest region clustered with the Southwest region. This contrasted with the clustered phonological distances also produced in Sanders (2008). In that experiment, there was no significant correlation between the inter-region phonological distances and syntactic distances.

There are several possible reasons for this lack of correlation. The two distance measures may find different dialect boundaries based on differences between syntax and phonology. Dialect boundaries may have shifted during the 40 years between the collection of the SED and the collection of the ICE-GB. One or both methods may be measuring the wrong thing. In this dissertation, although the focus remains on results of computational syntax distance as compared to traditional syntactic dialectology, the discussion compares recent phonological dialectometry results on Swedish to the results obtained here.

Nerbonne and Wiersma

Due to the lack of alignment between the larger corpora available for syntactic analysis, a statistical comparison of differences is more appropriate than the simple symbolic approach possible with the word-aligned corpora used in phonology. This statistical approach means that a syntactic distance measure will have to use counting as its basis.

Nerbonne and Wiersma (2006)'s method models syntax by part-of-speech (POS) trigrams and uses differences between trigram type counts in a permutation test of significance. The heart of the measure is simple: the difference in type counts between the combined features of two sites. Kessler (2001) originally proposed this measure, the RECURRENCE metric (R):

$$R = \sum_i |c_{ai} - c_{bi}| \tag{3.7}$$

Given two sites a and b, c_a and c_b are the feature counts. i ranges over all features, so c_{ai} and c_{bi} are the counts of sites a and b for feature i. R is designed to represent the amount of variation exhibited by the two sites while the contribution of individual features remains transparent to aid later analysis. Unfortunately, it doesn't indicate whether its results are significant; a permutation test is needed for that, described in section 3.4.

Dialectometry in British English

The methods used in this dissertation are an evolution of those in my previous work on British English: (Sanders 2007) and (Sanders 2008). There, I compared phonological and syntactic dialectometry as described above. The process is similar to Wiersma's work in (Nerbonne and Wiersma 2006) and (Wiersma 2009), but with variants of both feature set and distance measure.

The input is 30 interview sites (described in section 3.3). The sentences in each site have their features extracted (the features are described in section 3.2). Optionally, only 1000 sentences per site are sampled with replacement, but the site sizes, unlike the British interviews in my previous work, are fairly similar in size so this is only required for comparison to previous work. Then the features are counted, producing a mapping of feature types to token counts.

At this point, two sites are compared based on these feature counts. The feature counts are first normalized to account for variation in corpus size (described in the next section). Then they are converted to ratios, meaning that the counts are scaled relative to the other site. For example, counts of 10 and 30 would produce the ratio 1 to 3, as would the counts 100 and 300. Finally, the distance (described above in 3.2) is calculated 10 times and the result is averaged.

The sites are sampled by sentence rather than by feature because the intent is to capture syntax, where the composite unit is the sentence. Similarly, phonology's composite unit is the word—most processes operate within the word on individual segments; some processes operate between words but they are fewer. Therefore, the assumption that words are independent will lose some information but not the majority. In the same way, the basic unit of syntax is the sentence; processes operate on the words in the sentence, but inter-sentence processes are fewer. Because of this, the sites are sampled by sentence, combining the sentences of all speakers from an interview site.

This dissertation skips the per-speaker sampling of Wiersma's (2009) work on Finnish L2 speakers. I assume that, since discovery of dialect features is the goal of this research, the sentences of speakers from the same village are independent of the speaker, at least with respect to dialect features. Although the motivation is partly theoretical, there is also a difference between the Swediasyn dialect corpus, with 2–4 speakers for each of 30 sites, and Wiersma's L2 corpus, with dozens of speakers but only two groups. Sampling per-speaker would not be feasible for the Swediasyn because there aren't enough speakers per village.

Normalization

The two sites being compared can differ in size, even if the samples contain the same number of sentences; if one site contains many long sentences and the other contains many short ones, raw counts will favor the features extracted from the long sentences simply because each sentence yields more features. Additionally, the counts are converted to ratios to ignore the effect of frequency—in effect, this ranks features only by how much they differ between the two sites, ignoring the question of how often they occur relative to the other features extracted from the two sites. That is, a high ratio for a rare feature that happens only ten times in both sites is just as important as a high ratio

for a common feature that happens thousands of times.

The first normalization normalizes the counts for each feature within the pair of sites a and b. The purpose is to normalize the difference in sentence length, where longer sentences with more words cause features to be relatively more frequent than sites with many short sentences. Each feature count i in a vector, for example a, is converted to a frequency f_i

$$f_i = \frac{i}{N}$$

where N is the length of a. For two sites a and b this produces two frequency vectors, f_a and f_b.Then the original counts in a and b are redistributed according to the frequencies in f_a and f_b:

$$a'_i = \frac{f_{ai}(a_i + b_i)}{f_{ai} + f_{bi}}, b'_i = \frac{f_{bi}(a_i + b_i)}{f_{ai} + f_{bi}}$$

This redistributes the total of a pair from a and b based on their relative frequencies. In other words, the total for each feature remains the same:

$$a_i + b_i = a'_i + b'_i$$

but the values of a'_i and b'_i are scaled by their frequency within their respective vectors.

For example, assume that the two sites have 10 sentences each, with a site a with only 40 words and another, b, with 100 words. This results in $N_a = 40$ and $N_b = 100$. Assume also that there is a feature i that occurs in both: $a_i = 8$ and $b_i = 10$. This means that the relative frequencies are $f_{ai} = 8/40 = 0.2$ and $f_{bi} = 10/100 = 0.1$. The first normalization will redistribute the total count $(10 + 8 = 18)$ according to relative frequencies. So

$$a'_i = \frac{0.2(18)}{0.2 + 0.1} = 3.6/0.3 = 12$$

and

$$b'_i = \frac{0.1(18)}{0.2 + 0.1} = 1.8/0.3 = 6$$

Now that 8 has been scaled to 12 and 10 to 6, the fact that site b has more words has been normalized.

This reflects the intuition that something that occurs 8 of 40 times is more important than something that occurs 10 of 100 times.

The second normalization normalizes all values in both permutations with respect to each other. This is simple: find the average number of times each feature appears, then divide each scaled count by it. This produces numbers whose average is 1.0 and whose values are multiples of the amount that they are greater than the average. The average feature count is $N/2n$, where N is the number of feature occurrences and n is the number of feature types in the combined sites. Division by two is necessary since we are multiplying counts from a single permutation by summed counts from the combined sites' permutations. Each entry in the ratio vector now becomes

$$r_{ai} = \frac{2na'_i}{N}, r_{bi} = \frac{2nb'_i}{N}$$

For example, given the previous example numbers, this second normalization first finds the average. Assuming 5 unique features for a's 40 total features and 30 for b's total 100 features gives

$$n = 5 + 30 = 35$$

and

$$N = 40 + 100 = 140$$

Therefore, the average feature has $140/2(35) = 2$ occurrences in a and b respectively. Dividing $a'_i = 12$ and $b'_i = 6$ by this average gives $r_{ai} = 6$ and $r_{bi} = 3$. In other words, r_{ai} occurs 6 times more than the average feature.

Together, these normalizations control for the effect of variation in sentence length (the first normalization), corpus size (the second normalization), and relative overuse (the second normalization). Furthermore, the normalizations can be iterated, with the normalized output further renormalized. This exaggerates the differentiating effect of the normalization, which allows distance measures to be more sensitive to feature count variations.

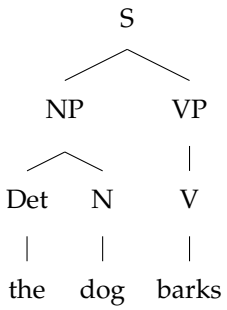
Syntax Features

In order to answer question 1, whether the distance measure agrees with dialectology, a distance measure such as R needs features that capture the dialect syntax of the interview corpus given as input. Following Nerbonne and Wiersma (2006), I start with parts of speech, then add the leaf-ancestor paths from my work on the ICE-GB (Sanders 2007), and finally add leaf-head paths and phrase-structure rules, as well as variants on these features. These feature sets each depend on a different type of automatic annotation, which is described in section 3.3.

Nerbonne and Wiersma (2006) argue that POS trigrams can accurately represent at least the important parts of syntax, similar to the way chunk parsing can capture the most important information about a sentence. If this is true, POS trigrams are a good starting point for a language model; they are simple and easy to obtain in a number of ways. They can either be generated by a tagger as Nerbonne and Wiersma did, or taken from the leaves of the trees of a syntactically annotated corpus as I did with the International Corpus of English (Sanders 2007).

Of course, bigrams are a possible feature since they are so similar to trigrams. I do not use them here for several reasons. First, previous work uses trigrams, so trigrams are needed in order to remain comparable. But bigrams offer only reduced sparseness and noise reductions compared to trigrams. However, neither feature sparseness nor noise is a problem for trigrams when used with the distance measures developed here, as will be seen in the results in chapter 4.

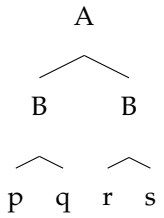
On the other hand, if syntax is in fact a phenomenon that involves hidden structure above the visible words of the sentence, a feature set should be constructed to capture that structure. Sampson's (2000) leaf-ancestor paths provide one way to do this: for each leaf in the parse tree, leaf-ancestor paths produce the path from that leaf back to the root. Generation is simple as long as every sibling is unique. For example, the parse tree



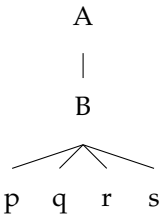
creates the following leaf-ancestor paths:

- S-NP-Det-the
- S-NP-N-dog
- S-VP-V-barks

For identical siblings, brackets must be inserted in the path to disambiguate the first sibling from the second. There is one path for each word, and the root appears in all four. However, there can be ambiguities if some node happens to have identical siblings. Sampson gives the example of the two trees



and



which would both produce

- A-B-p
- A-B-q
- A-B-r
- A-B-s

There is no way to tell from the paths which leaves belong to which B node in the first tree, and there is no way to tell the paths of the two trees apart despite their different structure. To avoid this ambiguity, Sampson uses a bracketing system; brackets are inserted at appropriate points to produce

- [A-B-p
- A-B]-q
- A-[B-r
- A]-B-s

and

- [A-B-p
- A-B-q
- A-B-r
- A]-B-s

Left and right brackets are inserted: at most one in every path. A left bracket is inserted in a path containing a leaf that is a leftmost sibling and a right bracket is inserted in a path containing a leaf that is a rightmost sibling. The bracket is inserted at the highest node for which the leaf is leftmost or rightmost.

It is a good exercise to derive the bracketing of the previous two trees in detail. In the first tree, with two B siblings, the first path is A-B-p. Since p is a leftmost child, a left bracket must be inserted, at the root in this case. The resulting path is [A-B-p. The next leaf, q, is rightmost, so a right bracket must be inserted. The highest node for which it is rightmost is B, because the rightmost leaf of A is s. The resulting path is A-B]-q. Contrast this with the path for q in the second tree; here q is not rightmost, so no bracket is inserted and the resulting path is A-B-q. r is in almost the same position as q, but reversed: it is the leftmost, and the right B is the highest node for which it is the leftmost, producing A-[B-r. Finally, since s is the rightmost leaf of the entire sentence, the right bracket appears after A: A]-B-s.

At this point, the alert reader will have noticed that both a left bracket and right bracket can be inserted for a leaf with no siblings since it is both leftmost and rightmost. That is, a path with two brackets on the same node could be produced: A-[B]-c. Because of this redundancy, single children are excluded by the bracket markup algorithm. There is still no ambiguity between two single leaves and a single node with two leaves because only the second case will receive brackets.

Sampson originally developed leaf-ancestor paths as an improved measure of similarity between gold-standard and machine-parsed trees, to be used in evaluating parsers. The underlying idea of a collection of features that capture distance between trees transfers quite nicely to this application. I replaced POS trigrams with leaf-ancestor paths for the ICE corpus and found improved results on smaller sites than Nerbonne and Wiersma had tested (Sanders 2007). The additional precision that leaf-ancestor paths provide appears to aid in attaining significant results.

Leaf-Head Paths

For dependency parses, it is easy to create a variant of leaf-ancestor paths called "leaf-head paths". Like leaf-ancestor paths, each word in the sentence is associated with a single leaf-head path. The difference is that the path is from the leaf to the head of the sentence via the intermediate heads. For example, the same sentence, "The dog barks", produces the following leaf-head paths, given the dependency parse in figure 3.1:

- root-V-N-Det-the

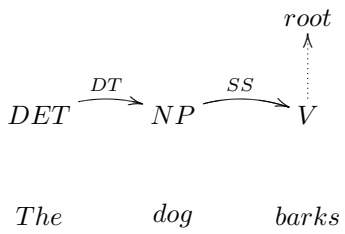


Figure 3.1: Dependency parse for "The dog barks."

- root-V-N-dog
- root-V-barks

The biggest difference between leaf-ancestor paths and leaf-head paths is the relative length of the paths: long leaf-ancestor paths indicate deep nesting of structure, while short ones indicate flatter structure. Length is a weaker indicator of deep structure for leaf-head paths; for example, the verb in a nested clause has a much shorter leaf-head path than leaf-ancestor path, but its dependents have comparable lengths between the two types of paths. Instead, length of path measures centrality to the sentence; longer leaf-head paths indicate less important words.

Leaf-head paths represent a compromise between leaf-ancestor paths and trigrams. Like trigrams, they capture lexical context, but the context is based on head dependencies, so long-distance context is possible. Like leaf-ancestor paths, they capture information about the nested structure of the sentence, although not as completely or explicitly.

Alternate Feature Sets

This section describes the variants besides the main feature sets already described above: trigrams, leaf-ancestor paths and leaf-head paths. Most are variants on these three main sets.

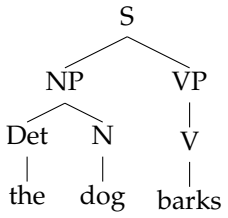


Figure 3.2: Example Tree

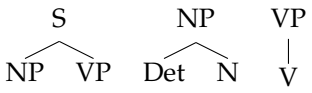


Figure 3.3: Phrase-Structure Rules Extracted

Part-of-Speech Unigrams

Part-of-speech unigrams are single parts of speech. Unlike POS trigrams, they do not capture context or order, only distributional differences. In this dissertation, they serve as a baseline since they are not expected to capture syntactic variation as much as the other feature sets.

Phrase Structure Rules

Phrase structure rules are extracted from the same parses as leaf-ancestor paths, but instead of capturing a series of parent-child relations, it captures single-level parent-child-sibling relations. For example, given the tree in figure 3.2 the extracted rules are given in figure 3.3.

Phrase structure rules are most similar to leaf-ancestor paths in emphasizing the hidden, parse structure of constituency parse trees. Unlike leaf-ancestor paths, they capture some context to the left and right. They also only cover one level in the tree, whereas leaf-ancestor paths traverse it from leaf to root. Phrase structure rules have the possibility to be useful in sentences where context is important, but they also depend on having accurate parses even at the top of the tree. This is difficult for automatic parsers to achieve.

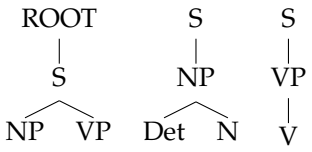


Figure 3.4: Grandparent Phrase-Structure Rules Extracted

Grandparent Phrase Structure Rules

Grandparent phrase structure rules are a variant of phrase structure rules that include the grandparent as well. Given the tree in figure 3.2, the extracted features are given in figure 3.4.

Grandparent phrase structure rules add some of the vertical information present in leaf-ancestor paths, hopefully without introducing data sparseness problems. However, they retain the advantage over leaf-ancestor paths of capturing left and right context.

Arc-Head Paths

As described in section 3.2, the usual labels for leaf-head paths are the leaves of the tree: 'root-V-N-Det-the' is the first leaf-head path for "The dog barks", which has the parts of speech "Det N V". However, one can also use the arc labels of the dependency parse to create arc-head paths. These paths have the same shape as their corresponding leaf-head paths, but use the labels of the dependency arcs between words instead of the parts of speech of the words themselves.

The sentence for the leaf-head example is given in figure 3.1, and the resulting arc-head paths are

- root-SS-DT-the
- root-SS-dog
- root-barks

Tags from Berkeley Parser

The Berkeley parser (Petrov et al. 2006), as described in section 3.3, can either tag incoming sentences with its own part of speech tagger, using the same splitting process as the rest of the parser, or with parts of speech specified externally. In this case, the external part-of-speech tagger is T'n'T (Brants 2000). Although the Berkeley-generated POS tags are not as good, it may be useful to see how they change the overall results—although it seems that accurate parts of speech are required for good features to be generated, it is useful to see how much the results degrade when given lower quality parts of speech. Note that the Berkeley-generated POS tags are used to generate trigrams and leaf-head paths, by taking the POS tags and feeding them into MaltParser.

To get the Berkeley parser to generate parts of speech, it is given the interview sites directly, skipping the tagging by T'n'T. Using the same method it uses to parsing the higher structure of the sentence, it also tags words with parts of speech. After parsing, these parts of speech are extracted from the leaves of the parse trees. First, the parts of speech are used to create trigram features. Second, the parts of speech are given to MaltParser for dependency parsing. This produces dependency parses based on parts of speech from the Berkeley parser. The result is trigrams and leaf-head paths based on Berkeley-generated POS tags instead of T'n'T-generated POS tags.

Dependencies from alternate MaltParser training

Since MaltParser uses Nivre's oracle-based dependency parsing algorithm, the default oracle, based on support vector machines, can be replaced with Timbl, the Tilburg Memory-Based Learner. It is possible that a memory-based learner improves parsing because support vector machines depend on large training corpora to provide good results. In contrast, a memory-based learner can obtain good results on limited training if the training happens to be representative and the right combination of parameters can be found for Timbl.

This is, however, somewhat complicated since Timbl is quite sensitive to parameter changes and usually requires specific tuning for particular tasks. To find the best parameters, I use a manual search across a number of the major distance measures provided by Timbl, as well as fallback-combinations from more complicated distance measures to less complicated ones.

Each combination was evaluated with ten-fold cross-validation on Talbanken. The best combination was Jeffrey divergence with 5 nearest neighbors, no feature weighting, inverse distance neighbor weighting, and fallback to the Overlap metric for fewer than two neighbors. Jeffrey divergence is a symmetric variant of Kullback-Leibler divergence, also described in section 3.2. These parameter settings were used as a basis for parsing and generation of leaf-ancestor paths.

Combining Feature Sets

Combining feature sets gives the classifier more information about a site by combining the information that each feature set captures. This dissertation uses a simple linear combination. In other words, all features are counted together with equal weight. This is easy and should allow the feature ranker to find a greater variety of features that capture the same underlying syntactic information.

Alternate Distance Measures

There are several reasons to test distance measures besides R. There are a couple of a priori reasons for this: R is fairly simple, so more complicated variations on it may provide better sensitivity at the expense of sensitivity to noise. Also, variations explore the measure space better in case that R is not significant for some combination of corpus/feature set.

Post-hoc, there are interesting patterns of statistical significance produced by the combination of distance measure and feature set. These patterns are not trivially obvious. This is not expected, but may provide insight into the measure/feature combination, which helps resolving Hypotheses 1 and 2.

Kullback-Leibler divergence

Kullback-Leibler divergence, or relative entropy, is described in Manning and Schütze (1999). Relative entropy is similar to R but more widely used in computational linguistics. The name relative

entropy implies an intuitive interpretation: it is the number of bits of entropy incurred when compressing a site b with the optimal compression scheme for a second site a. Unless the two sites are identical, the relative entropy $KL(a||b)$ is non-zero because a's optimal compression scheme will over-compress b's features that are more common in a than in b, whereas it will under-compress features that are less common in a than in b.

For example, assume that site a has two features with type counts {S-NP-N : 20, S-VP-PP-N : 10}. An optimal compression scheme for a would compress S-NP-N twice as much as S-VP-PP-N because it occurs twice as often. However, if this compression scheme is used on a site b with the feature counts {S-NP-N : 15, S-VP-PP-N : 15}, efficiency will be worse; S-NP-N and S-VP-PP-N occur the same number of times in b, so the smaller compressed size of S-NP-N will be used less often than expected, while the larger compressed size of S-VP-PP-N will be used more. This difference can be measured precisely for each feature:

$$a_i \log \frac{a_i}{b_i}$$

where a_i is type count of the ith feature in a and b_i is the type count of the ith feature in b. This measures the number of bits lost, or entropy, for each feature i. Like R's differences, the per-feature entropy can be summed to find the total entropy. In the example above, the entropy for S-NP-N is $20 \log \frac{20}{15} = 5.75$.

However, Kullback-Leibler divergence as defined is a divergence: it measures the divergence of features in the site b from the features of site a. A dissimilarity is required for dialectology, which means that the divergence must additionally be symmetric. A divergence can be made symmetric by calculating it twice: the divergence from a to b added to the one from b to a. The complete formula is given in equation 3.8 and the complete example is worked in equation 3.9.

$$KL(a||b) = \sum_i a_i \log \frac{a_i}{b_i} + b_i \log \frac{b_i}{a_i} \tag{3.8}$$

$$(20 \log \frac{20}{15} + 15 \log \frac{15}{20}) + (10 \log \frac{10}{15} + 15 \log \frac{15}{10}) = (5.75 - 4.32) + (-4.05 + 6.08) = 3.46 \quad (3.9)$$

Jensen-Shannon divergence

Several variants of relative entropy exist that lift various restrictions from the input distributions. One is Jensen-Shannon divergence (Lin 1991), which was designed as a dissimilarity from the start. It uses the same denominator for both directions: the average of the two frequencies. That means that each feature's entropy is found using the following formula:

$$a_i \log \frac{b_i}{(a_i + b_i)/2} + b_i \log \frac{a_i}{(a_i + b_i)/2}$$

There is a common subexpression in this value: $(a_i + b_i)/2$: the average of the two features. If we let $\bar{c_i} = (a_i + b_i)/2$ rewrite the formula to take advantage of this simplification, we get equation 3.10.

$$JS = \sum_i a_i \log \frac{b_i}{\bar{c_i}} + a_i \log \frac{b_i}{\bar{c_i}} \quad (3.10)$$

Unlike Kullback-Leibler divergence, Jensen-Shannon divergence does not require that features exist in both sites being compared in order to be counted. KL divergence cannot count unique features, in fact, because if either a_i or b_i is zero, then it will divide by zero at some point. The current implementation of KL divergence simply skips zero values, which means it ignores features unique to a particular site. Jensen-Shannon divergence avoids this problem because it divides by $\bar{c_i}$, the average of the feature counts. Because KL divergence ignores features unique to one site, it should be less susceptible to noise than JS divergence. This can be useful in the presence of unreliable annotators. However, KL divergence will also produce less detail in the case that unique features are useful.

Cosine similarity

Cosine similarity is used in many parts of computational linguistics and related areas such as information extraction and data mining. Nerbonne and Wiersma (2006) use it as reference point for comparison to previous work in these areas. Cosine similarity measures the similarity between two high-dimensional points in space. Each feature is modeled as a dimension, and the type count from each site is plotted as a point on that dimension. In equation 3.11, vectors a and b are multiplied, then divided by the product of their lengths. This equation can be written in an element-by-element way as in equation 3.12. Here, the vector multiplication is written out as the sum of pairwise products. The vector length of a and b is written as a square root of sum of squares.

$$\frac{a \cdot b}{||a||||b||} \tag{3.11}$$

$$\frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2} \cdot \sqrt{\sum_i b_i^2}} \tag{3.12}$$

Interestingly, the results for this measure are somewhat different from the other distance measures, possibly because, unlike the others described, it is not a linear sum.

3.3 Input Processing

To investigate the first question, agreement with dialectology, I need a dialect corpus that can be syntactically annotated (3.3); if it is not already annotated, it must be possible to annotate it automatically so I can avoid time-consuming manual annotation. Automatic annotation requires a syntactically annotated training corpus (3.3) and parsers for the models of syntax I use as a basis (3.3).

SweDiaSyn

In order to find dialect distance between sites, a dialect corpus that can be syntactically annotated is required. The dialect corpus used in this dissertation is SweDiaSyn, the Swedish part of the ScanDiaSyn. SweDiaSyn is a transcription of SweDia 2000 (Bruce et al. 1999). SweDia 2000 was collected between 1998 and 2000 from 97 locations in Sweden and 10 in Finland. Each location has 12 interviewees: three 30-minute interviews for each of older male, older female, younger male and younger female. However, the SweDiaSyn transcriptions do not yet include all of SweDia 2000; the completed transcriptions currently focus on older speakers.

Currently there are 36,713 sentences of transcribed speech from 49 sites, an average of 749 sentences per site. However, the sites range from 110 to 1780 sentences because some sites have fewer complete transcriptions than others.

In the SweDiaSyn, there are two types of transcription: standard Swedish orthography, with glosses for words not in standard Swedish; and a phonetic transcription for dialects that differ greatly from standard Swedish. For this dissertation, the orthographic/gloss transcription is used so that lexical items are comparable across dialects. However, only 30 of the 49 sites have been glossed, so the total usable size of the corpus is 21,004 sentences, with an average of 700 sentences per site. The sites range from 301 to 1,144 sentences.

Talbanken

Because SweDiaSyn consists of unannotated lexical items only, Talbanken05, a syntactically-annotated corpus, is used to train the automatic annotators which in turn annotate SweDiaSyn. Talbanken05 is a treebank of written and transcribed spoken Swedish, roughly 300,000 words in size. It is an updated version of Talbanken76 (Nivre et al. 2006b); Talbanken76's trees are annotated following a scheme called MAMBA; Talbanken05 adds phrase structure annotation and dependency annotation using the annotation formats TIGER-XML and Malt-XML. In addition to syntactic annotation, Talbanken is lexically annotated for morphology and part-of-speech.

Parsing

In order to build the language models described above, SweDiaSyn must be POS tagged, constituency parsed and dependency parsed. This allows the features to be extracted for use by the distance measure.

Before annotation, the SweDiaSyn sentences are cleaned in order to improve the output of the parsers. Cleaning the sentences consists of removing restarts, stops, and mumbled words, which are all marked in the transcription.

Tags 'n' Trigrams

The Tags 'n' Trigrams (T'n'T) tagger (Brants 2000) is used for tagging, with the POS annotations from Talbanken05 used as training. T'n'T is an efficient Markov-model trigram tagger, meaning that it uses only the previous two words to decide on the part of speech for a word. It backs off to even less context in the case of sparse data; if the trigram composed of the current word and the previous two words has not been seen before, most of the decision will be based on the current word and one previous word. T'n'T handles unknown words by a simple form of suffix analysis—unknown words that have similar endings to known words are more likely to get that tag.

Berkeley Parser

For constituency parsing, the Berkeley parser (Petrov et al. 2006) is trained on Talbanken05. The Berkeley parser has shown good performance on languages other than English, which is not common for constituency parsers (Petrov and Klein 2007).

The Berkeley parser learns latent annotations, which means that it learns grammars from training data by assuming that the training gives a coarser picture than the true grammar, but one that is also too specialized to the observed sentences. For example, it may start with the single category NP for noun phrases, which is too coarse to capture the subject/object distinction. So the category will be split into NP1 and NP2. However, because the splits are cutoff to a certain level of frequency, not every random characteristic of the training data is learned.

MaltParser

For dependency parsing, MaltParser will be used with the existing Swedish model trained on Talbanken05 by Hall, Nilsson and Nivre. Dependency parsing proceeds similarly to constituency parsing; the dependency structures of Talbanken05 are cleaned and normalized, then used to train a parser.

MaltParser is an inductive dependency parser that uses a machine learning algorithm to guide the parser at choice points (Nivre et al. 2006a). This means that the parsing algorithm is deterministic. Although this sounds impossible for an ambiguous language, it achieves this by relying on a machine learner to choose the correct option at points where multiple options exist. The machine learner is either a memory-based learner or a support vector machine trained on a history-based model. This model uses the internal state of the parser as features for training.

3.4 Output Analysis

After a distance measure has been defined for the interview sites within the dialect corpus (see above, section 3.2) and syntactic features have been extracted (3.2), the results must be tested for significance (3.4). The significant results must be analyzed by clustering (3.4) and multi-dimensional scaling (3.4) to determine which dialect regions are found by the distance measure. Finally, the most highly ranked features used to produce the dialect distances must be enumerated (3.4).

Permutation test

To find out if a dialect distance value is significant as measured, a permutation test with a Monte Carlo technique described by Good (1995) is required, following closely the same usage by Nerbonne and Wiersma (2006). The intuition behind the technique is to compare the distance between two sites with the distance between two random subsets of the shuffled, combined sites. This shuffling and subsets is repeated multiple times. If the random subsets' distance is less than the distance of the two actual sites more than p percent of the time, then we can reject the null

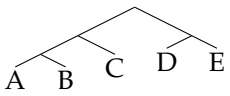


Figure 3.5: Hierarchical Cluster Dendrogram

hypothesis that the two sites were are actually drawn from the same dialect: that is, we can assume that the two sites are different. The reason is that the distance should be larger between the samples of the original corpora than the distance between the random subsets: any real differences will be randomly redistributed between both subsets by the shuffling process.

To see how this works, for example, assume that the distance between the two British regions London and Scotland is some value such as 100. The permutation test then shuffles London and Scotland to create a combined site "LSMixed" and splits it into two random sites. Since the real differences between London and Scotland are now mixed between the two random subset, we would expect the distance between these two subsets to be less than 100. This should be true at least 95% of the time for the distance 100 to be significant according to the normal threshold of significance, $p < 0.05$, in the social sciences.

Cluster Analysis

The first question, agreement with dialectology, requires a clustering method to allow inter-site distances to be compared more easily. The dendrogram that binary hierarchical clustering produces allows easy visual comparison of the most similar sites. An example is given in figure 3.5.

A clustering algorithm provides understanding of which sites group together. There are a variety of clustering algorithms, but hierarchical clustering is the most appropriate method for this problem because it does not specify the number of groups ahead of time. Other clustering algorithms such as k-means or expectation maximization require the number of expected clusters to be given. Hierarchical clustering creates its clusters implicitly by grouping items using a binary merge operation. The merge is repeated until a tree with a single root is formed. The implicit clusters can be extracted by looking at which speakers share the same subtree, as well as looking for large differences in internal node heights connecting subtrees.

The initial step for any clustering algorithm is to find distances between all pairs of sites as described above. These distances between all pairs of sites result in a set of high-dimensional spatial relationships. While they could be analyzed as such, such high-dimensional distances are difficult to visualize. The job of a clustering algorithm is to reduce the dimensionality and create a useful visualization of the relative positions of the speakers. Hierarchical clustering does this by creating a tree—if there are similarities between the speakers, it should be obvious by looking at the tree.

There are some complications, however. Because the clustering algorithm is nothing but repeated merges, it is not always clear at what level the best clusters are formed. For example, Clopper and Pisoni (2004) used a similar clustering algorithm on perceptual dialect data from American English speakers and found two distinct North/South clusters for most features. These two clusters had less defined sub-clusters as well: the Western speakers of American English usually grouped with the North cluster but slightly separated from the other dialects in the North cluster. Of course as the sub-clusters become smaller, they usually become less distinctive because the distances are smaller. Ultimately, human judgment is necessary to determine what is a cluster and what is not.

Bottom-up hierarchical clustering

With hierarchical clustering defined, "bottom-up" now needs a definition. As in any tree-building problem, the two obvious ways one can build a tree are top-down and bottom-up. Bottom-up clustering works in the following way. First, each site is put into its own group. Then the algorithm determines the distance between each pair of groups. The closest two groups are merged into a single group and the process is repeated until a single root group is created. This process is bottom-up because it creates the terminal nodes of the tree first and builds up the internal structure of the cluster tree from there.

For example, figures 3.6 through 3.10 show the sequence of merges need to produce the dendrogram in figure 3.5. On the first step A and B are merged (figure 3.7), followed by D and E (figure 3.8). Then C merges with the A-B cluster (figure 3.9). Finally the A-B-C cluster and the D-E cluster merge to form a single tree (figure 3.10).

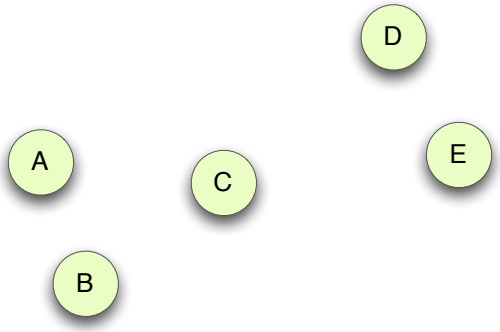


Figure 3.6: Sites Before Clustering

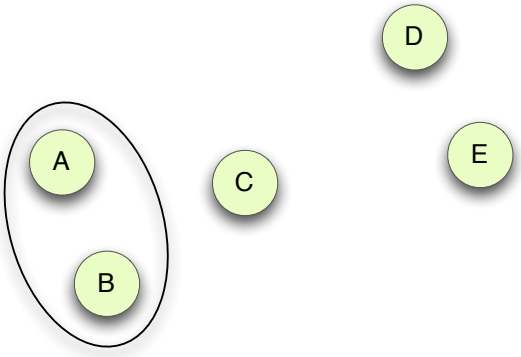


Figure 3.7: Sites After A-B Merge

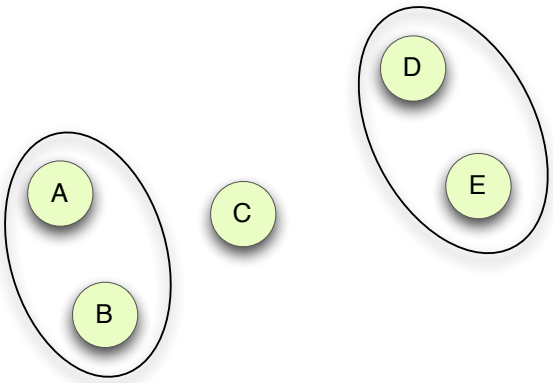


Figure 3.8: Sites After D-E Merge

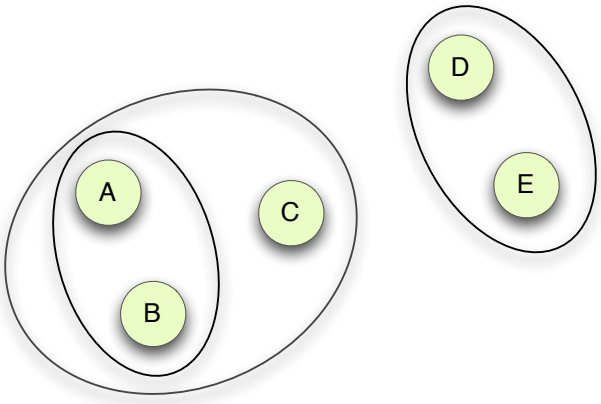


Figure 3.9: Sites After A-B-C Merge

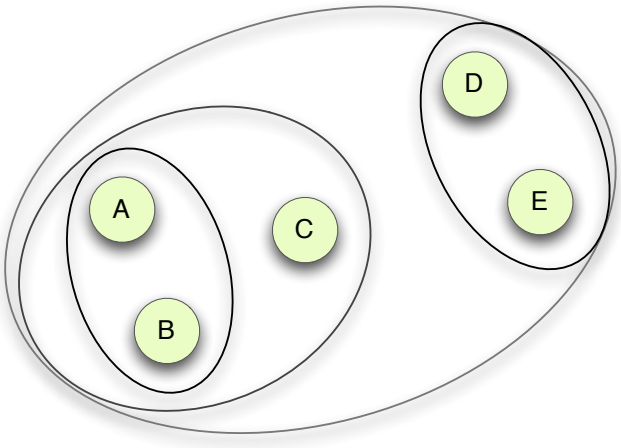


Figure 3.10: Sites After Clustering

	B	C	D	E
A	10	20	40	50
B		20	50	40
C			30	30
D				12

Table 3.1: Example dissimilarities

To find the two closest groups, Ward's (1963) method is used. At each merge step, this method evaluates every possible binary merge. Each merge is given a score that minimizes some objective function—here, the average of distances between sites in the new group. The best merge replaces its children and the process repeats until a singly rooted tree is created. For n sites, this takes $n - 1$ iterations, because at each step, two groups are merged.

For example, using the distances in table 3.1 for the five example sites, the first merge is trivial since each site starts in its own singleton tree (figure 3.11): the distance between A and B, 10, is the minimum and thus the best. This produces the forest in figure 3.12.

The distances between the A-B tree and the others are now more complicated to calculate: the A-B-C merge has a distance of $10 + 20 + 20/3$ = 16.6. This is smaller than the A-B-D merge ($10 +$

Figure 3.11: Ward's method, before clustering

Figure 3.12: Ward's method, after A-B merge

$40 + 50/3 = 33.3$), but larger than the D-E merge ($12/1 = 12$) which eventually turns out to be the smallest merge, producing figure 3.13.

The next merge is primarily concerned with where C will merge, whether with A-B or D-E; an A-B-D-E merge is much larger at $10 + 40 + 50 + 50 + 40 + 12/6 = 33.6$. As previously calculated, the A-B-C merge is 16.6, while a C-D-E merge is $30 + 30 + 12/3 = 24$. So the new merge is A-B-C, producing figure 3.14.

The two remaining trees are merged. Here, the final value of the objective function is the average of all distances in the table, that is $302/10 = 30.2$. The final tree is given in figure 3.15.

Ward's method is less efficient than other common clustering methods, but it usually finds small, round clusters, making it worth the extra computer time. In contrast, single-link distance, for example, compares only the two closest elements of the two members of a possible merge. This is faster, but is susceptible to creating thin, oval groups—even though the bulk of a group may be distant, a single outlier usually leads to a bad grouping, which recursively leads to further outliers.

Consensus Trees

A weakness of cluster dendrograms is that small variations in distances can cause large changes in the cluster membership of sites. Consensus trees circumvent this weakness by combining the results of multiple related dendrograms. Only the clusters that occur in the majority of dendrograms appear in the consensus tree. For a survey of consensus trees, see chapter 6 of (Bryant 1997).

Figure 3.13: Ward's method, after D-E merge

Figure 3.14: Ward's method, after A-B-C merge

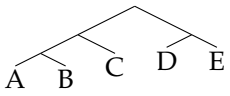


Figure 3.15: Ward's method, after clustering

Consensus trees can be constructed by an algorithm with three primary steps. First, the algorithm finds the spans of every internal node in every tree. That is, each non-terminal node N is replaced with the terminal nodes $w_i \ldots w_j$ that it dominates. Second, the algorithm counts span types and retains only the spans that occur in a majority of dendrograms. For example, if there are 9 dendrograms, the consensus tree will contain only spans that occur in 5 or more of them. Third, the spans are reconstructed into a single tree. Nodes that no longer have a direct parent are added to a higher ancestor.

For example, consider the three hierarchical dendrograms of figure 3.16, adapted from the example given by Amenta et al. (2003). They cluster the input set {A B C D E} three different ways. The majority-rule consensus tree for these three trees is given in figure 3.17.

The spans for the internal nodes of the three trees are given in figure 3.18. There is quite a bit of overlap at higher levels, but near the leaves, {B C} and {C D} vary, as do {B C D} and {C D E}. As a result, when the spans are combined and counted, {B C} and {C D E} only appear once. This means that they will be dropped from the consensus tree because they appear in 1/3 of the trees and because 1/3 is less than or equal to 1/2, they are not majority spans.

Reconstruction is fairly simple; taken from the top down, both {A} and {B C D E} must be children of {A B C D E}. Because {A} is not a member of the majority spans, it is added directly as

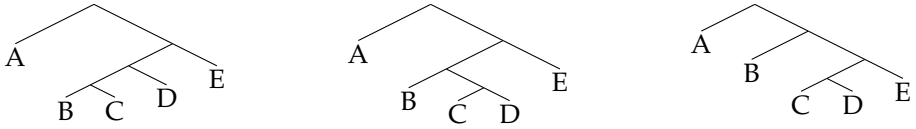


Figure 3.16: Input cluster dendrograms

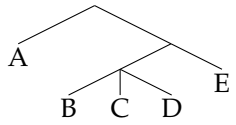


Figure 3.17: Output consensus dendrogram

{B C}	{C D}	{C D}
{B C D}	{B C D}	{C D E}
{B C D E}	{B C D E}	{B C D E}
{A B C D E}	{A B C D E}	{A B C D E}

Figure 3.18: Spans from input trees

{B C}	1 / 3	*
{C D}	2 / 3	
{B C D}	2 / 3	
{C D E}	1 / 3	*
{B C D E}	3 / 3	
{A B C D E}	3 / 3	

Figure 3.19: Span type frequencies (starred rows do not occur in the majority of trees)

{C D}
{B C D}
{B C D E}
{A B C D E}

Figure 3.20: Majority span types

a child of {A B C D E}. Although this occurs in all three original trees, this is not the case when {B C D} adds {B} as a child. The result is that {B C D} has ternary branching, which was not present in any of the original trees.

As can be seen from the example, high degrees of branching in the consensus tree near the leaves indicate that the original trees do not agree well. Therefore, it is not safe to draw conclusions from an original tree in the areas where it disagrees with other original trees.

Composite Clustering

A visual alternative to a consensus tree is composite, or fuzzy, clustering (Kleiweg et al. 2004). Instead of removing clusters that do not appear in the majority of cluster trees, fuzzy clustering plots every cluster tree on a map completely. However, each tree is plotted transparently. If a large number of trees agree on a cluster, the cluster will be plotted many times, creating a dark line. Conversely, clusters without wide agreement will only get a light line. This provides a graphical equivalent to consensus trees.

To make this work, hierarchical clustering must change slightly: the input is a diagonal matrix of distances as before, but the output is no longer a binary tree but a diagonal matrix of distances, like the input. The distances between two sites are now the number of clusters that separate them.

You can then draw this directly on a map: put a line equidistant between each pair of sites, making it darker the more clusters that separate them.

But this still relies on hierarchical clustering, which is not very stable. To work around this, I use multiple hierarchical cluster trees based on the parameter variations described above in section 3.2. Each set of parameters generates one dendrogram, which produces a matrix of cluster-separation counts for each parameter variant in feature set, distance measure, normalization count, and sampling method. I average cluster-separation count matrices and scale the line darkness accordingly to get a more stable picture of which boundaries are important.

	B	C	D	E
A	10	20	40	50
B		20	50	40
C			30	30
D				12

Table 3.2: Example dissimilarities

Multi-dimensional scaling

Multi-dimensional scaling (MDS) is an alternate approach to making the high-dimensional dissimilarities more easily interpretable. Instead of creating a tree, multi-dimensional scaling reduces the high-dimensional dissimilarities to 3 dimensions, which can then be represented using (Red,Green,Blue) color triples. When painted on a map, these colors provide a nice visualization of the regions that similar sites form as well as how sharp the boundaries are with other regions.

MDS of dissimilarities uses Kruskal's (1964)a method to reduce m dissimilarities to an n dimensional space, where $n < m$. Since dissimilarities do not satisfy the triangle inequality, the mapping to lower-dimensional space will require some change of the dissimilarities. The question is how to minimize the change. Kruskal's method defines a global measure of how much change is required called Stress. An initial Stress is obtained by sorting the dissimilarities by size and measuring how far each dissimilarity would have to change in order for all the dissimilarities to be mapped to points in n-dimensional space. This initial stress is reduced by a process of gradient descent as described in (Kruskal 1964b), which ultimately results in the minimum overall distortion of the dissimilarities.

Correlation

Correlation is useful on two levels. First, correlation between dialect distance and other measures provides an idea of how well dialect distance agrees with these other measures. Although traditional dialectology measures do not provide numeric distances to correlate with syntactic dialect distance, other distances, such as geographic distance, travel distance, and phonological dialect distance do. These correlations provide circumstantial evidence that the distances are valid. Second,

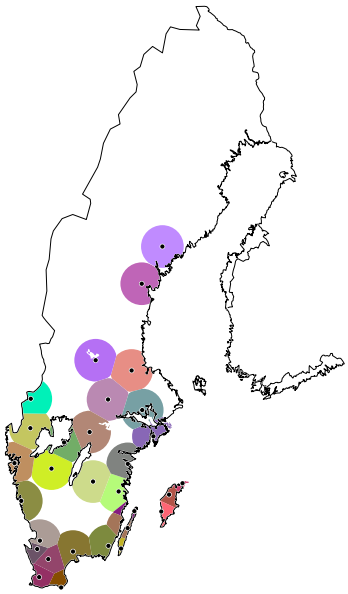


Figure 3.21: Swedia, Multi-Dimensional Scaling of Trigrams measured by Jensen-Shannon divergence

it is interesting to measure correlation between varying combinations of feature set and distance measure. Disparate combinations may end up by producing distances that correlate highly.

For geographic and travel distance, distances were obtained from Bing Maps. Travel distance is not guaranteed to be completely accurate, but even so should correlate more highly with dialect distance than straight-line geographic distance. Still, travel distance was defined in terms of Sweden's roads and ferries in 2010. This may differ from the travel distances over the time period for which dialect usage was strongest. Gooskens (2004) estimates Norwegian travel times for 1900; a similar estimate for Sweden would probably give even higher correlation.

$$r_{ab} = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2}\sqrt{\sum_{i=1}^{n}(b_i - \bar{b})^2}} \tag{3.13}$$

Correlation uses Pearson's r, which is defined in equation 3.13 (Aron et al. 2006). Pearson's r takes two vectors, called a and b here. In the equation, $\bar{a}$ and $\bar{b}$ are the averages of a and b respectively. Like distance itself, correlation must be checked for significance. Mantel's test provides tests Pearson's r for significance between inter-connected sites (Mantel 1967). Mantel's test is much like the permutation test for significance described above. It first finds the correlation between two sets of distances. One distance result set is permuted repeatedly and at each step correlated with the other set. The original correlation is significant if the permuted correlation is lower than the original correlation more than 95% of the time.

Feature Ranking

Feature ranking is needed to compare distances qualitatively to the Swedish dialectology literature; the most important features should be similar to those discussed most by dialectologists when comparing regions. Without feature ranking of some kind, there is no way to relate the quantitative distances between sites with the features that contribute most to those distances.

A simple feature ranking for distance is easy for one-to-one site comparisons; each feature's normalized weight is equal to its importance in determining the distance between the two sites. See figure 3.22: features that appear more often in one site of the compared pair are negative, while

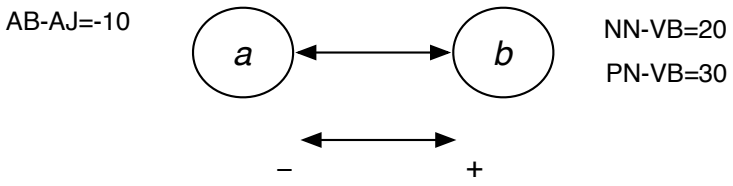


Figure 3.22: Feature-ranking 1:1

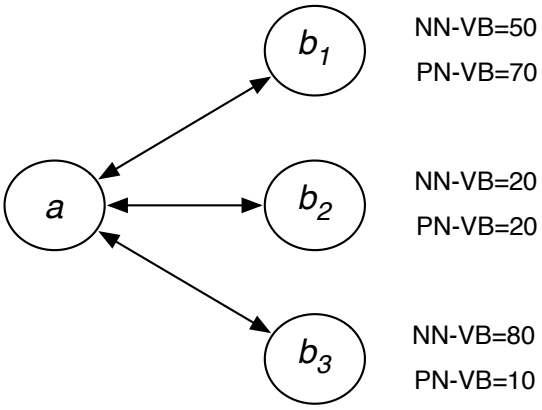


Figure 3.23: Feature-ranking 1:Many

features that appear more often in the other site are positive. In the example, the bigram AB-AJ occurs more often in the left-hand-site, while NN-VB and PN-VB occur more often in the right-hand-site. This is the same as the first step of R and R^2: R then takes the absolute value of this difference and R^2 takes the square.

Features can be ranked between a single site and multiple site by averaging. For example, in figure 3.23, the binary comparison between the left-hand site and each of the three right-hand sites produces three sets of features. The features can be combined by averaging the score for each feature type. NN-VB's averaged score would be $50 + 20 + 80/3 = 50$, for example.

This average can be extended to compare two sets of sites. As can be seen in figure 3.24, each feature on the right-hand side is no longer a single number, but an average of the comparison against each sites on the left-hand side. Therefore, in this example, NN-VB's overall average score is

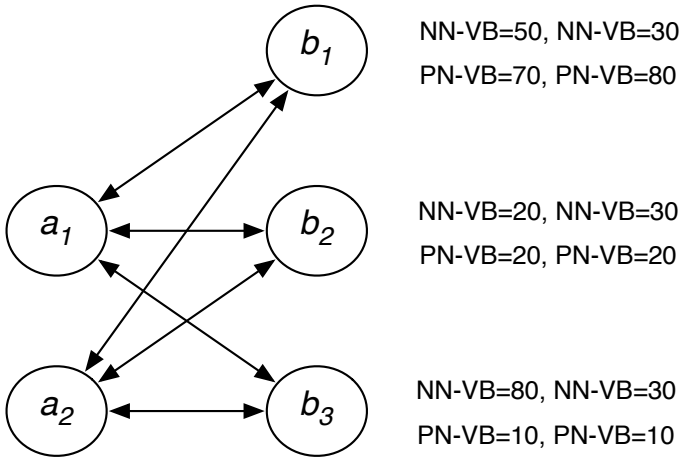


Figure 3.24: Feature-ranking Many:Many

$$\left(\frac{50+30}{2}+\frac{20+30}{2}+\frac{80+30}{2}\right)/3=\frac{50+30+20+30+80+30}{6}=40$$

Wiersma's Measure of Feature Overuse

Wiersma (2009) uses a method similar to the one described above, but with an additional normalization intended to show which features are used relatively more in one or the other of the two sites to be compared. This normalization is similar to the second step of the normalization described in section 3.2: it also removes the effect of frequency. However, this normalization removes the effect of frequency from the difference of the two sites, whereas the second removes it from the distance between two sites. Both normalizations are similar in being of limited value for the noisier data, automatically annotated. However, the results could improve if the data are sufficiently clean.

The normalization centers the two counts around 1.0 by dividing each count by the average count, scaled by the total size of the two sites. The equation for a feature i with paired counts a_i and b_i is given in equation 3.14 for site a. There, N_a and N_b are the sizes of the sites a and b, and N is the combined size of the two sites.

$$o_{ai} = a_i / \frac{(a_i + b_i)N_a}{N_a + N_b} \tag{3.14}$$

The equation can be simplified slightly; equations of this form for both sites a and b are given in equation 3.15.

$$o_{ai} = \frac{a_i(N_a + N_b)}{(a_i + b_i)N_a} \text{ and } o_{bi} = \frac{b_i(N_a + N_b)}{(a_i + b_i)N_b} \tag{3.15}$$

For example, the previous one-to-one example shows that NN-VB occurs 20 more times in site b than in a. Let this arise from the counts $a_{\text{NN-VB}} = 10$ and $b_{\text{NN-VB}} = 30$. Also let the site sizes for a and b be $N_a = 40$ and $N_b = 100$ respectively. For the overuse-normalized numbers, this gives

$$o_a\text{NN-VB} = \frac{10(100 + 40)}{(10 + 30)40} = 1400/1600 = 0.875$$

and

$$o_b\text{NN-VB} = \frac{30(100 + 40)}{(10 + 30)100} = 4200/4000 = 1.05$$

With the frequencies turned into ratios of overuse, it is possible to see that the additional 20 occurrences in site b are not that important; they only give o_bNN-VB a value of 1.05. Unlike frequencies, identical overuses occur between pairs of the same ratio rather than the same difference. In other words, $a_i = 2, b_i = 12$ gives the same overuse, $o_b = 1.2$, as $a_j = 10, b_j = 60$, because in both cases $b = 6a$. In contrast, frequency comparison gives a difference of $b_i - a_i = 10$ and $b_j - a_j = 50$, making the second difference much more important.

3.5 Conclusion

Besides giving an overview of the previous work, this chapter covered the methods used in this dissertation, which fall into three categories: distance measure methods, methods for processing dialect corpora to be used with a distance measure, and methods for processing distances to be compared with dialectology results. The last two, while not the focus of the research, are complex,

partly because of the dissimilarity between the required inputs and outputs and partly because dialectology presents its results in many forms, meaning that many methods are required to produce comparable forms from dialectometry distances. The results, in chapter 4, mirror the order of the last two sections, with the bulk devoted to the output processing that produces maps that are comparable to dialectology.

4

Results

These results are meant to answer two main questions: first, how well does this approach to syntactic dialectometry agree with dialectology? Second, what combinations of distance measures, feature sets and other settings produce the best results for linguistic analyses? Additionally, the results are meant to allow comparison with phonological dialectometry.

The organization of this chapter mirrors the order of the methods chapter, particularly the output analysis (section 3.4). First, there is an overview of the different parameter settings, the combinations of distance measure and feature set, as well as other settings (section 4.1). Then the number of significant distances for each parameter setting is given (section 4.2), which is followed by the correlation with geography and travel distance for each parameter setting (section 4.3). These sections focus mainly on detecting which settings do not produce valid results, so that they can be ignored in the rest of the chapter. At a high level, they answer the question of the suitability of statistical syntactic dialectometry: whether or not significant results can be found.

Next, the specific dialectological results are examined. First, cluster dendrograms provide a visualization of which sites the distance measures find to be similar (section 4.4). In addition, to improve the reliability of the dendrograms, consensus trees (section 4.4) and composite cluster maps are produced (section 4.4). Next, multi-dimensional scaling gives a smoother view of similarity than clusters (section 4.5). Finally, features are ranked and extracted from each cluster in the consensus tree (section 4.6).

Feature Set
Leaf-Ancestor Path
Part-of-speech Trigram
Leaf-Head Path
Phrase Structure Rule
PSR with Grandparent
Part-of-speech Unigram
Leaf-Head Path, based on Timbl training
Leaf-Arc Path
All features combined

Measure
R
R^2
Kullback-Leibler divergence
Jensen-Shannon divergence
cosine dissimilarity

Sampling Method
1000 sentences
All sentences

Iterations of normalization
1
5

Table 4.1: Settings for the five parameters tested

4.1 Parameter Settings

There are 180 parameter settings investigated in this chapter. This number arises from the four parameters: measure, feature set, sampling method and number of normalization iterations. 5 measures, 9 feature sets, 2 sampling methods and 2 iterations of normalization gives $5 \times 9 \times 2 \times 2 = 180$ different settings. The settings are given in table 4.1.

In addition, the size of each of the 30 interview sites are given in table 4.2.

4.2 Significant Distances

Significant distances help answer the question whether a syntactic measure has succeeded in finding reliable distances; the measure will always return some distance, but if the sites are too small, it may not be significant. Therefore the results should have few non-significant distances. In the tables, the total number of comparisons between all 30 sites is the $435 = 30(30-1)/2$. In the first set, tables 4.3 – 4.4, the results are shown from one iteration of the normalization step. In the second set, tables 4.5 – 4.6, the results from five normalization iterations are shown.

Bold numbers in the tables indicate that fewer than 95% of the distances were significant. In

Site	Sentences	Words	Site	Sentences	Words
Ankarsrum	630	7708	Leksand	923	10676
Anundsjo	1144	11897	Loderup	429	7850
Arsunda	937	8933	Norra Rorum	546	9160
Asby	693	7171	Orust	1067	11409
Bara	696	10724	Ossjo	481	12275
Bengtsfors	663	7423	Segerstad	837	9746
Boda	1029	17425	Skinnskatteberg	730	9529
Bredsatra	360	6938	Sorunda	768	11144
Faro	659	8260	Sproge	381	4399
Floby	557	6392	StAnna	876	13156
Fole	727	9920	Torsås	374	9217
Frillesas	572	9634	Torso	956	15577
Indal	1126	13090	Vaxtorp	903	11353
Jamshog	301	8661	Viby	431	6734
Köla	528	10133	Villberga	680	11479

Table 4.2: Size of Interview Sites

	R	R^2	KL	JS	cos
Leaf-Ancestor	0	0	11	0	0
Trigram	0	0	0	0	0
Leaf-Head	0	0	0	0	0
Phrase-Structure Rules	0	0	**95**	0	0
Phrase-Structure with Grandparents	0	0	**273**	0	0
Unigram	0	0	0	0	0
Leaf-Head with MaltParser trained by Timbl	0	0	**47**	0	0
Leaf-Arc Labels	0	0	0	0	0
All Features Combined	0	0	0	0	0

Table 4.3: Number of non-significant distances for sample size 1000, 1 normalization

table 4.6, the 5-iteration table that compares full sites, the only combination with *less* than 5% non-significant results is cosine dissimilarity with unigram features, marked with italics. Note that here, 5% is an arbitrary cutoff point not related to the usual significance cutoff $p < 0.05$; the basis for these tables are themselves number of significant distances found.

Analysis of the significance of dialect distance provides a measure of how reliable the distances to be analyzed later in this chapter are. A distance that does not find significant distances between the 30 sites is not suitable for precise inspection, although small numbers of non-significant distances will still allow methods to return interpretable results.

The highest number of significant distances are found in the first case (table 4.3): 1 round of normalization with a fixed-size sample of 1000 sentences. From there, both full-site comparisons

	R	R^2	KL	JS	cos
Leaf-Ancestor	7	11	12	**35**	9
Trigram	4	1	0	**24**	1
Leaf-Head	10	12	20	**44**	19
Phrase-Structure Rules	**26**	17	**24**	**49**	20
Phrase-Structure with Grandparents	**58**	**35**	**38**	**71**	**33**
Unigram	1	2	0	0	2
Leaf-Head with MaltParser trained by Timbl	11	21	18	**74**	**30**
Leaf-Arc Labels	14	19	**37**	**94**	17
All Features Combined	0	0	1	8	2

Table 4.4: Number of non-significant distances for complete sites, 1 normalization

	R	R^2	KL	JS	cos
Leaf-Ancestor	5	**56**	**34**	0	0
Trigram	3	2	0	0	0
Leaf-Head	3	14	4	0	0
Phrase-Structure Rules	11	4	**66**	1	0
Phrase-Structure with Grandparents	18	0	**109**	4	0
Unigram	**52**	**53**	15	17	0
Leaf-Head with MaltParser trained by Timbl	7	20	**45**	0	0
Leaf-Arc Labels	6	**54**	17	1	0
All Features Combined	0	4	0	0	0

Table 4.5: Number of non-significant distances for sample size 1000, 5 normalizations

	R	R^2	KL	JS	cos
Leaf-Ancestor	**290**	**284**	**287**	**278**	**204**
Trigram	**284**	**283**	**283**	**276**	**196**
Leaf-Head	**293**	**286**	**285**	**279**	**211**
Phrase-Structure Rules	**289**	**294**	**286**	**275**	**236**
Phrase-Structure with Grandparents	**285**	**290**	**286**	**270**	**258**
Unigram	**297**	**296**	**294**	**293**	*9*
Leaf-Head with MaltParser trained by Timbl	**294**	**289**	**288**	**284**	**222**
Leaf-Arc Labels	**294**	**290**	**291**	**293**	**162**
All Features Combined	**279**	**279**	**279**	**269**	**191**

Table 4.6: Number of non-significant distances for complete sites, 5 normalizations

(table 4.4) and 5 rounds of normalization (table 4.5) have fewer significant distances, although the number is still usable. However, the combination of the two, with 5 rounds of normalization over full-site comparisons, has only one combination with fewer than 5% of distances that are *not* significant. Although both full-site comparisons and multiple rounds of normalization may increase the precision of the results, their combined effect on significance is so detrimental that its results are useless. For the rest of the analysis, the combination of full-site comparison and 5 rounds of normalization will be skipped.

Significance by Measure

The distance measures most likely to find significance are, in order, cosine dissimilarity, Jensen-Shannon divergence and R. Each method had different parameter settings for which it was stronger. For 1000-sentence sampling, tables 4.3 and 4.5, cosine similarity resulted in all significant distances, even for part-of-speech unigrams, which are intended as the baseline feature set. Excluding unigrams, Jensen-Shannon divergence has similar performance. For full-site comparisons, tables 4.4 and 4.6, both perform considerably worse; surprisingly, both perform better on unigram features, Jensen-Shannon so much so that it is the only feature set for which it finds all significant distances. R, on the other hand, performs decently on all combinations of parameter settings; its low significance for phrase structure rules is shared by Kullback-Leibler and Jensen-Shannon divergences.

When comparing the performance of Kullback-Leibler and Jensen-Shannon divergence it is not surprising that Jensen-Shannon outperforms Kullback-Leibler on fixed-size sampling. Although both are called "divergence", Jensen-Shannon divergence is actually a dissimilarity. Recall that the divergence from point A to B may differ from the divergence from point B to A. A divergence like Kullback-Leibler can be converted to a dissimilarity by measuring $KL(A, B)+KL(B, A)$. However, this dissimilarity must skip features unique to a single site in order to avoid division by zero. This means that for smaller sites Kullback-Leibler loses information that Jensen-Shannon is able to use. On the other hand, while this may explain Kullback-Leibler's improved performance for full-site comparisons, it doesn't explain Jensen-Shannon's much worse performance.

Significance by Feature Set

For 1 round of normalization, the best feature sets are the simple ones: trigrams and unigrams, as well all combined features. On the other hand, trigrams and leaf-head paths (with its variations) are the best feature sets with 5 rounds of normalization. However, the variation isn't strong; any feature set can give good results with the right distance measure. The problem is that no clear patterns emerge.

The relatively high quality of trigrams and unigrams does not make sense given only the linguistic facts; however, it is likely that the entirely automatic annotation used here introduces more and more errors as more annotators run, operating on previous automatic annotations. Trigrams are the result of only one automatic annotation, and one for which the state of the art is near human performance. So the fact that these particular parts of speech are of higher quality than the corresponding dependencies or constituencies is probably the deciding factor in their higher number of significant distances.

Given the above facts, the question should rather be: why do leaf-head paths perform as well as they do? Better, for example, than the leaf-ancestor paths on which they're modeled: why does more normalization hurt leaf-ancestor paths but not leaf-head paths? It could be that there is less room for error; many of the common leaf-head paths are short: short interview sentences with simple structure make for shorter leaf-head paths than leaf-ancestor paths. As a result, the important leaf-head paths consist mainly of a couple of parts-of-speech. This difference in feature length holds for any length of sentence, but is exaggerated for simple sentences, where the amount of structure generated for a phrase-structure parse for a clause is more than for a dependency parse. In general, clauses, embedded and otherwise, produce the largest difference in amount of structure between the two, so the feature length differs for deeply nested sentences as well.

Another reason could be a difference in parsers: MaltParser has been tested on Swedish by its designers (Nivre et al. 2006a). Besides English, the Berkeley parser has been tested prominently on German and Chinese. Therefore, the difference would better be explained by appealing to the difference in parsers rather than an unsuitability of Swedish for constituent analysis.

It is disappointing linguistically that trigrams provide the most reliable results so far; a linguist

would expect that including syntactic information would make it easier to measure the differences between sites. If it is, as hypothesized here, an effect of chaining machine annotators, a study using a manually annotated corpus could detect this. However, it still means that trigrams are the most useful feature set from a practical view, because automatic trigram tagging is very close to human performance with little training. That means the only required human work is the transcription of interviews in most cases.

On the other hand, if additional features sets are to be developed for a corpus, then combining all available features seems to be a successful strategy. The distance measures seem to be able to use all available information for finding significant distances.

4.3 Correlation

In dialectology, the default expectation for dialect distance is that it correlates with geographic distance (Chambers and Trudgill 1998). A lack of correlation does not necessarily mean that a measure is invalid, but presence of correlation means that the distance measure substantiates the well-known tendency of dialect distributions to be more or less smoothly gradient over physical space.

In addition, distance measures are more likely to correlate significantly with travel distance than with straight-line geographic distance. This makes sense since the difficulty of moving from place to place is what influences dialect formation, and taking roads into account is an improved estimate over straight-line distance.

The tables that present geographic and travel correlation, 4.7 – 4.14, mark significant correlations with a star for $p < 0.05$, two stars for $p < 0.01$ and three stars for $p < 0.001$. However, these correlations are only trustworthy in the case that the underlying distances are significant. Significant correlations from significant distances (as cross-referenced from tables 4.3 – 4.6) are marked by italics.

Besides this, correlation between combinations of measure/feature set can show how closely related they are–in other words, how similarly they view the underlying data which remains the

	R	R^2	KL	JS	cos
Leaf-Ancestor	-0.01	0.03	0.02	-0.02	0.08
Trigram	0.17	0.17	0.10	0.19	0.13
Leaf-Head	-0.06	0.03	0.00	-0.07	0.05
Phrase-Structure Rules	0.01	*0.18**	0.16	0.01	0.12
Phrase-Structure with Grandparents	0.03	*0.25**	0.21*	0.03	0.12
Unigram	*0.18**	0.17	*0.29***	*0.30***	*0.18**
Dependencies, MaltParser trained by Timbl	-0.07	0.02	-0.00	-0.08	0.05
Arc-Head	-0.07	0.06	-0.06	-0.09	0.00
All Features Combined	-0.02	0.03	0.01	-0.02	0.07

Table 4.7: Geographic correlation for sample size 1000, 1 normalization iteration

	R	R^2	KL	JS	cos
Leaf-Ancestor	0.02	0.09	0.11	-0.00	0.09
Trigram	*0.27**	*0.26**	*0.30***	0.21*	0.08
Leaf-Head	-0.03	0.12	0.14	-0.06	0.02
Phrase-Structure Rules	0.13	*0.36***	0.30**	0.11	*0.20**
Phrase-Structure with Grandparents	0.15	0.41**	0.36**	0.14	0.19*
Unigram	*0.20**	*0.20**	*0.33***	*0.33***	*0.22**
Dependencies, MaltParser trained by Timbl	-0.02	0.14	0.16	-0.05	0.02
Arc-Head	-0.06	0.13	-0.01	-0.12	-0.03
All Features Combined	0.03	0.11	0.16	-0.00	0.04

Table 4.8: Geographic correlation for complete sites, 1 normalization iteration

same for all. It is analyzed in section 4.3.

This is similar to the reasoning behind correlation with geography—but the assumption is that geography is a factor underlying dialect formation; while the distance measure measures some aspect of the language which we hope is dialects, it is indirectly (even less directly) measuring the geography. Therefore, correlation with geography should occur.

Third, correlation with corpus size is not predicted and is probably an undesired defect in sampling or normalization. Correlation with corpus size is presented in tables 4.16 – 4.19.

From tables 4.7 – 4.14 we see that parameter settings that correlate significantly do so at rates around 0.2 to 0.3, with a high of 0.37 for phrase-structure-rule features measured by R^2, 1 normalization iteration and comparison of full sites. The significant correlations are mostly concentrated in the trigram, unigram and combined feature sets.

	R	R^2	KL	JS	cos
Leaf-Ancestor	0.14	0.14	0.16	0.15	0.08
Trigram	*0.22**	0.17	*0.22**	*0.22**	0.16
Leaf-Head	0.10	0.11	0.15	0.12	0.10
Phrase-Structure Rules	0.14	0.10	0.14	0.15	0.06
Phrase-Structure with Grandparents	0.16	0.14	0.14	0.15	0.05
Unigram	0.12	0.11	0.14	0.13	0.17
Dependencies, MaltParser trained by Timbl	0.09	0.12	0.16	0.11	0.11
Arc-Head	0.08	0.10	0.14	0.10	0.09
All Features Combined	0.19	0.16	*0.20**	*0.21**	0.11

Table 4.9: Geographic correlation for sample size 1000, 5 normalizations

	R	R^2	KL	JS	cos
Leaf-Ancestor	-0.14	-0.16	-0.15	-0.15	-0.08
Trigram	-0.09	-0.07	-0.09	-0.09	-0.09
Leaf-Head	-0.22	-0.21	-0.18	-0.22	-0.10
Phrase-Structure Rules	-0.19	-0.14	-0.11	-0.20	-0.01
Phrase-Structure with Grandparents	-0.17	-0.11	-0.09	-0.18	-0.02
Unigram	-0.10	-0.06	-0.07	-0.08	0.14
Dependencies, MaltParser trained by Timbl	-0.19	-0.18	-0.18	-0.19	-0.10
Arc-Head	-0.21	-0.18	-0.18	-0.21	-0.10
All Features Combined	-0.18	-0.18	-0.16	-0.18	-0.09

Table 4.10: Geographic correlation for complete sites, 5 normalizations

	R	R^2	KL	JS	cos
Leaf-Ancestor	-0.03	0.02	0.01	-0.04	0.07
Trigram	0.20	0.19	0.11	*0.23**	0.14
Leaf-Head	-0.07	0.01	-0.01	-0.08	0.05
Phrase-Structure Rules	0.01	*0.18**	0.17	0.00	0.14
Phrase-Structure with Grandparents	0.03	*0.26**	0.22*	0.03	0.15
Unigram	*0.20**	*0.19**	*0.30***	*0.31***	*0.21**
Dependencies, MaltParser trained by Timbl	-0.08	0.02	-0.01	-0.09	0.05
Arc-Head	-0.08	0.05	-0.06	-0.10	0.00
All Features Combined	-0.03	0.03	0.01	-0.03	0.06

Table 4.11: Travel correlation for sample size 1000, 1 normalization iteration

	R	R^2	KL	JS	cos
Leaf-Ancestor	0.02	0.08	0.11	0.00	0.08
Trigram	*0.31**	*0.28**	*0.32***	0.26*	0.09
Leaf-Head	-0.02	0.12	0.13	-0.05	0.01
Phrase-Structure Rules	0.15	*0.37***	0.32**	0.13	*0.22**
Phrase-Structure with Grandparents	0.17	0.43**	0.38**	0.16	0.22*
Unigram	*0.22**	*0.22**	*0.33***	*0.34***	*0.24**
Dependencies, MaltParser trained by Timbl	-0.01	0.14	0.17	-0.04	0.02
Arc-Head	-0.06	0.12	-0.02	-0.12	-0.03
All Features Combined	0.04	0.10	0.16	0.01	0.04

Table 4.12: Travel correlation for complete sites, 1 normalization iteration

	R	R^2	KL	JS	cos
Leaf-Ancestor	0.17	0.19*	0.17*	0.18	0.07
Trigram	*0.24**	*0.20**	*0.25**	*0.26**	0.16
Leaf-Head	0.14	0.16	0.17	0.15	0.10
Phrase-Structure Rules	0.17	0.14	0.16*	0.18	0.06
Phrase-Structure with Grandparents	0.19	*0.18**	0.17*	0.19	0.06
Unigram	0.15	0.13	*0.17**	0.16	*0.20**
Dependencies, MaltParser trained by Timbl	0.12	0.16	0.18	0.14	0.11
Arc-Head	0.09	0.13	0.14	0.11	0.08
All Features Combined	*0.23**	*0.20**	*0.22**	*0.24**	0.11

Table 4.13: Travel correlation for sample size 1000, 5 normalizations

	R	R^2	KL	JS	cos
Leaf-Ancestor	-0.13	-0.13	-0.10	-0.13	-0.04
Trigram	-0.06	-0.04	-0.05	-0.06	-0.05
Leaf-Head	-0.20	-0.17	-0.13	-0.19	-0.06
Phrase-Structure Rules	-0.15	-0.08	-0.05	-0.15	0.04
Phrase-Structure with Grandparents	-0.12	-0.05	-0.03	-0.13	0.03
Unigram	-0.07	-0.03	-0.04	-0.05	*0.18**
Dependencies, MaltParser trained by Timbl	-0.18	-0.15	-0.12	-0.18	-0.05
Arc-Head	-0.20	-0.17	-0.14	-0.19	-0.06
All Features Combined	-0.16	-0.14	-0.11	-0.15	-0.05

Table 4.14: Travel correlation for complete sites, 5 normalizations

Analysis

As with the number of significant distances, trigrams and unigrams are the most likely to correlate with geographic and travel distance, as well as the combined feature set for the 5-normalization parameter setting. Note that in tables 4.7 – 4.14, the significant correlations are marked with an asterisk, but only the italicized correlations are based on at least 95% significant distances. For example, this means that most of the significant correlations based on phrase-structure rules are not valid.

It is worthwhile to note, however, that the valid and significant correlations based on phrase-structure grammars give the highest correlations: 0.37 for R^2 with full-site comparisons and 1 round of normalization. The addition of more data and more normalization is interesting in expanding the correlating parameter settings beyond those that include unigram features. It may be that this is an instance of the noise/quality tradeoff. These additions appear to extract more detail from the data, at the cost of additional interference from noisy data.

Inter-measure Correlation

Correlation between measures shows that they produce similar results. It also suggests that they use similar information to do so. For example, cosine similarity correlates the least with the others, which means that its results are the least like the others. It also implies that cosine similarity uses information from input features differently than the other measures. Since the performance of the summed, non-cosine measures is a little better for this site size, practical use of this distance method should probably start with them. In other computational linguistic applications, cosine distance is typically used with larger corpora, so it is possible that it provides better results with larger corpora, such as corpora based on entire provinces of Sweden rather than the individual villages used in this dissertation.

The average correlation between different measures is given in table 4.15. The correlations are averaged over the correlations for all combinations of feature set with 1000-sentence samples and with non-significant correlations removed before averaging.

	R^2	KL	JS	cos
R	0.85	0.85	0.98	0.39
R^2		0.90	0.83	0.57
KL			0.88	0.67
JS				0.44

Table 4.15: Average Inter-measure-correlation of measures

	R	R^2	KL	JS	cos
Leaf-Ancestor	-0.38	-0.26	-0.37	-0.40	-0.37
Trigram	0.12	-0.12	-0.16	0.14	-0.18
Leaf-Head	-0.39	-0.26	-0.35	-0.43	-0.39
Phrase-Structure Rules	0.06	0.15	0.00	0.03	-0.10
Phrase-Structure with Grandparents	0.08	0.19	0.07	0.04	-0.09
Unigram	-0.08	-0.14	-0.09	-0.09	-0.10
Dependencies, MaltParser trained by Timbl	-0.35	-0.23	-0.28	-0.37	-0.37
Arc-Head	-0.44	-0.26	-0.40	-0.48	-0.34
All Features Combined	-0.37	-0.26	-0.38	-0.42	-0.40

Table 4.16: Size correlation for sample size 1000, 1 normalization

The inter-measure correlation is essentially a summary of the results from the significance testing and correlations. R and Jensen-Shannon produce nearly identical results, and also correlate highly. Cosine similarity is quite different from the other measures, though the correlation is still higher than with travel distance. This is expected insofar as the cosine operation at the heart of cosine similarity differs more from the sums of absolute values or logarithms of other measures.

Correlation with Corpus Size

As previously stated, correlation with corpus size is not predicted and is probably an undesired defect in sampling or normalization. Correlation with corpus size is presented in tables 4.16 – 4.19.

Corpus size between two sites can be measured in two different ways: either by the sum of the sites' sizes, or by the difference. Here the sum is used: a larger sum means more tokens. If there is a correlation with size, it must arise because higher token counts are not properly normalized. In other words, two large sites will have more tokens, leading to higher type counts, which directly leads to higher distances. Smaller sites will lead to lower distances.

In tables 4.16 and 4.17, the 1-normalized correlations, only two correlations are significant.

	R	R^2	KL	JS	cos
Leaf-Ancestor	-0.19	-0.15	-0.16	-0.24	-0.36
Trigram	*0.30**	0.08	0.19	0.08	-0.39
Leaf-Head	-0.17	-0.06	-0.08	-0.26	-0.41
Phrase-Structure Rules	0.52**	*0.40***	0.30*	0.47**	-0.21
Phrase-Structure with Grandparents	0.54**	0.43**	0.37**	0.50**	-0.22
Unigram	-0.09	-0.13	-0.11	-0.13	-0.13
Dependencies, MaltParser trained by Timbl	-0.08	0.02	0.09	-0.14	-0.39
Arc-Head	-0.32	-0.16	-0.26	-0.40	-0.35
All Features Combined	-0.15	-0.11	-0.10	-0.25	-0.42

Table 4.17: Size correlation for complete sites, 1 normalization

	R	R^2	KL	JS	cos
Leaf-Ancestor	*0.35**	0.36**	0.06	0.27	-0.32
Trigram	*0.75***	*0.63***	*0.46***	*0.68***	-0.24
Leaf-Head	*0.46***	*0.44***	0.14	*0.38***	-0.33
Phrase-Structure Rules	*0.85***	*0.59***	0.36**	*0.85***	-0.34
Phrase-Structure with Grandparents	*0.88***	*0.66***	0.40**	*0.88***	-0.36
Unigram	0.38**	0.35**	0.14	0.19	-0.04
Dependencies, MaltParser trained by Timbl	*0.44***	*0.41***	0.16	*0.39**	-0.30
Arc-Head	0.20	0.28*	-0.00	0.09	-0.28
All Features Combined	*0.58***	*0.48***	0.21	*0.47***	-0.31

Table 4.18: Size correlation for sample size 1000, 5 normalizations

	R	R^2	KL	JS	cos
Leaf-Ancestor	-0.55	-0.38	-0.26	-0.53	-0.17
Trigram	-0.29	-0.27	-0.19	-0.26	-0.14
Leaf-Head	-0.61	-0.43	-0.27	-0.58	-0.18
Phrase-Structure Rules	-0.21	-0.08	-0.04	-0.22	-0.14
Phrase-Structure with Grandparents	-0.24	-0.08	-0.03	-0.26	-0.14
Unigram	-0.38	-0.25	-0.30	-0.32	-0.08
Dependencies, MaltParser trained by Timbl	-0.52	-0.33	-0.20	-0.51	-0.15
Arc-Head	-0.59	-0.45	-0.33	-0.54	-0.20
All Features Combined	-0.61	-0.44	-0.26	-0.55	-0.18

Table 4.19: Size correlation for complete sites, 5 normalizations

However, in table 4.18, 5-normalized correlations with 1000-sampling, a large number of correlations are significant. Specifically, the highest performing measures, R, R^2, and Jensen-Shannon divergence, correlate significantly with size for nearly all feature sets. Since this is not a predicted correlation, it means that these distances may be invalid. However, another piece of evidence makes this conclusion uncertain: geographic distance also correlates with corpus size at a rate of 0.31, $p < 0.01$, and travel distance correlates at 0.32, $p < 0.01$. This correlation is also unexpected, since there is no reason to expect that distance predicts corpus size or vice versa. However, it shows that the size correlation of dialect distance may at least by partly explained here by the unexpected correlation with geographic and travel distance. Therefore, 5-normalized results will be presented throughout the rest of the results.

Analysis

The correlation of corpus size and dialect distance is a problem. It is not a predicted as a side effect of the way dialect distance is measured. The fact that travel distance also correlates with corpus size at a rate of 0.32 confuses the issue further. Is corpus size the determining variable? Or is there an unknown variable influencing all three? One possibility is "interviewer boundaries", common in corpora collected by multiple people (Nerbonne and Kleiweg 2003). Perhaps a single interviewer improved with practice and collected longer interviews as the interview collection progressed. Or perhaps cultural differences between the interviewer and interviewees caused some participants in one area to talk more than in another area.

Although the size correlation of the dialect distances may be explained by the correlation with geographic/travel distance, they are still somewhat worrying. There is a great enough difference above the correlation of corpus size and geographic/travel distance that 5-normalized distances might not be reliable. However, if 5-normalization introduces a dependency on corpus size, then the distances from full-corpus comparisons should correlate even more highly. This is not the case.

Alternatively, it is possible that the fixed-size sampling method is not properly eliminating size differences between interview sites. Future work should develop a method for normalizing a comparison between two full sites. It should avoid sampling, but also take the relative number of

sentences into account.

4.4 Clusters

Cluster dendrograms provide a visualization of which sites the distance measures find most similar. They are formed in a bottom-up manner, repeatedly merging the two most similar groups at each step until only one group remains. The resulting dendrogram usually has obvious subtrees which can be treated as clusters. By grouping sites into clusters, cluster dendrograms allow closer comparison to dialectology than correlation. These clusters can be compared to the regions proposed by syntactic dialectology.

The first two dendrograms in this section hold feature set, measure, and sample size constant at trigrams, Jensen-Shannon, and 1000-sentence samples, respectively. Then they vary the amount of normalization: figure 4.1 has 1 normalization round, while figure 4.2 has 5. These two examples were chosen because of their high numbers of significant distances and correlation with travel distance; the highest correlation of 5-normalized distances with travel distance, 0.26, is with the Jensen-Shannon measure and trigram features in figure 4.2.

The third figure, figure 4.3, gives the dendrogram for the parameter settings with the highest travel distance correlation, phrase-structure rules, 1 normalization, 1000-sentence samples, and R^2 measure. The highest correlation of 1-normalized distances with travel distance, 0.37, is given by R^2 measured over phrase-structure-rule features, comparing full sites. Those parameter settings produce the dendrogram in figure 4.3.

Unlike the significances, cosine similarity's dendrograms are fairly similar to those of other features. See for example figure 4.4, with cosine, trigram features and 5 iterations of normalization.

However, it is difficult to judge the amount of agreement between these individual dendrograms. These figures are mostly given as examples rather than for in-depth comparison. Instead of manually comparing each to the dialect regions of Sweden, a better option is to aggregate them automatically into a single dendrogram, retaining only the clusters that agree. This is a consensus tree.

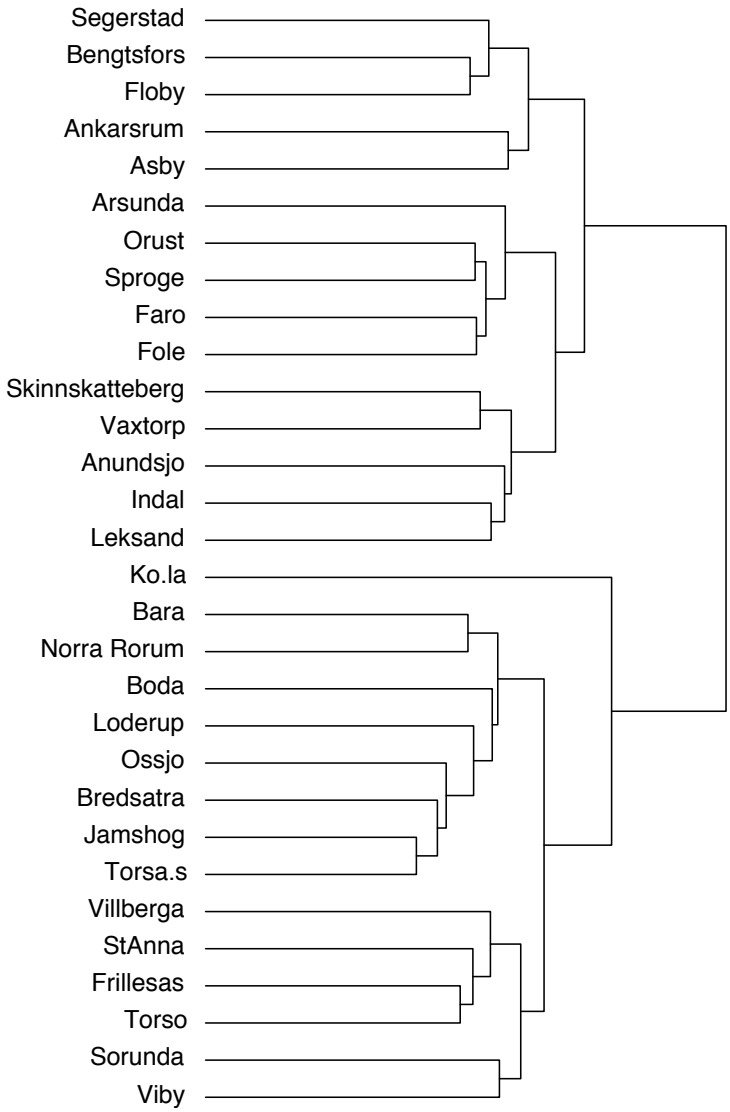


Figure 4.1: Dendrogram With Jensen-Shannon measure and trigram features, 1 normalization, 1000 samples

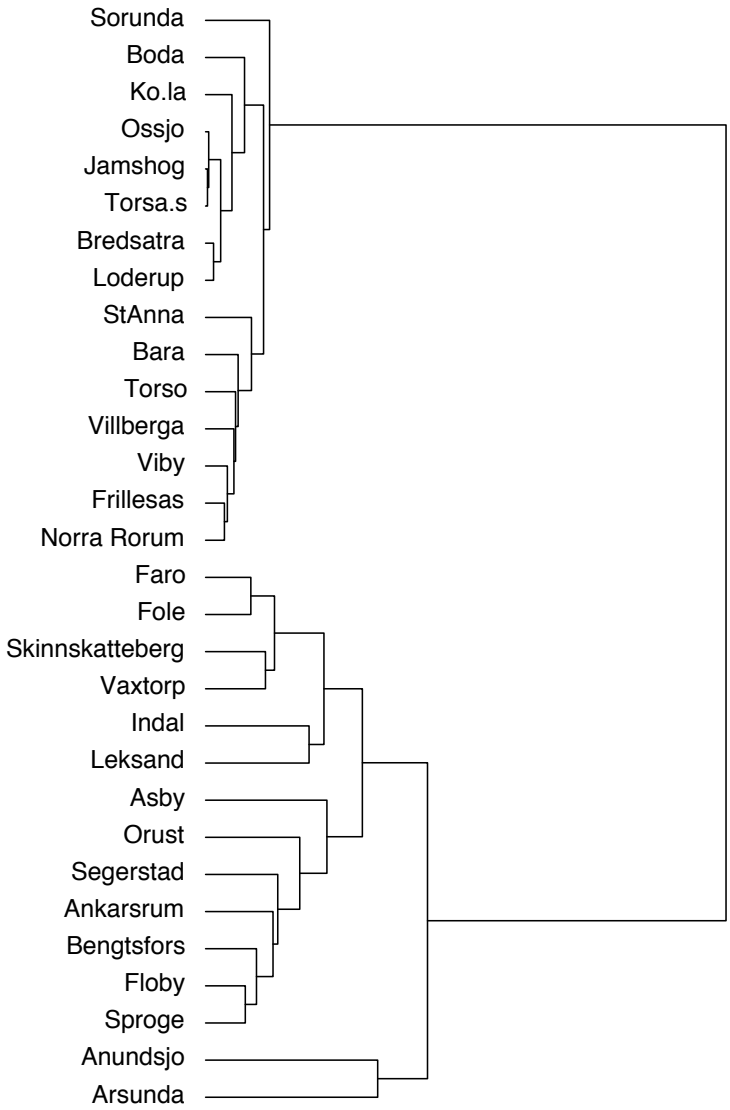


Figure 4.2: Dendrogram With Jensen-Shannon measure and trigram features, 5 normalizations, 1000 samples

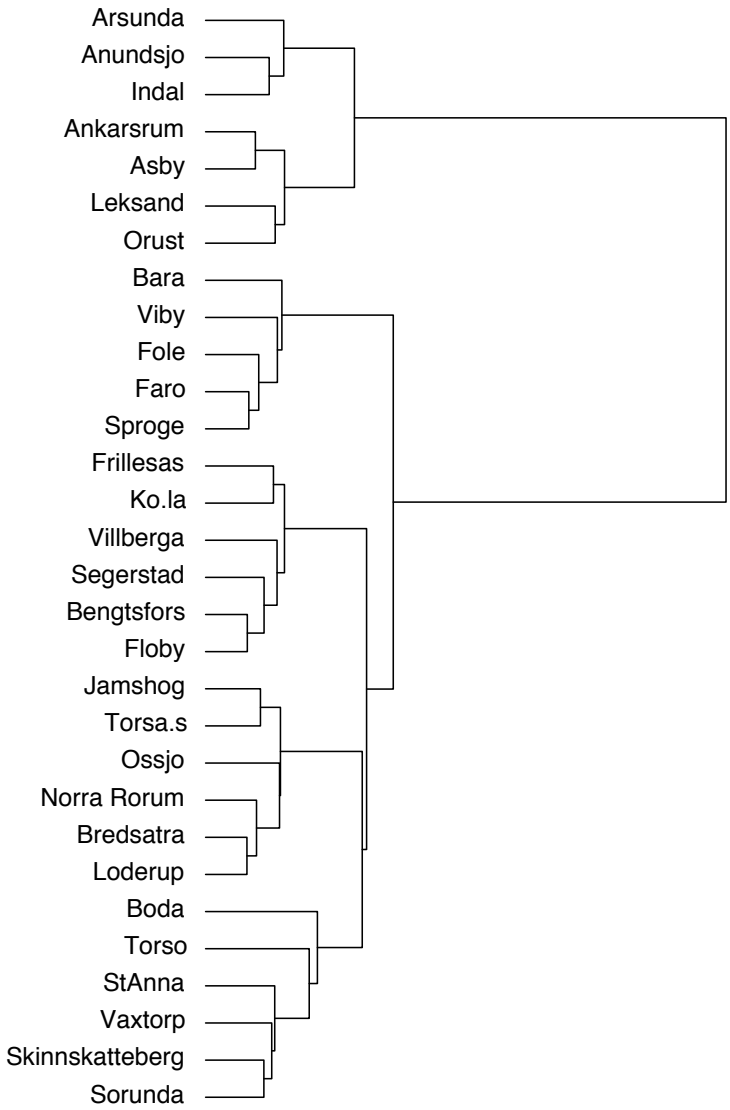


Figure 4.3: Dendrogram With R^2 measure and phrase-structure-rule features, 1 normalization, complete sites

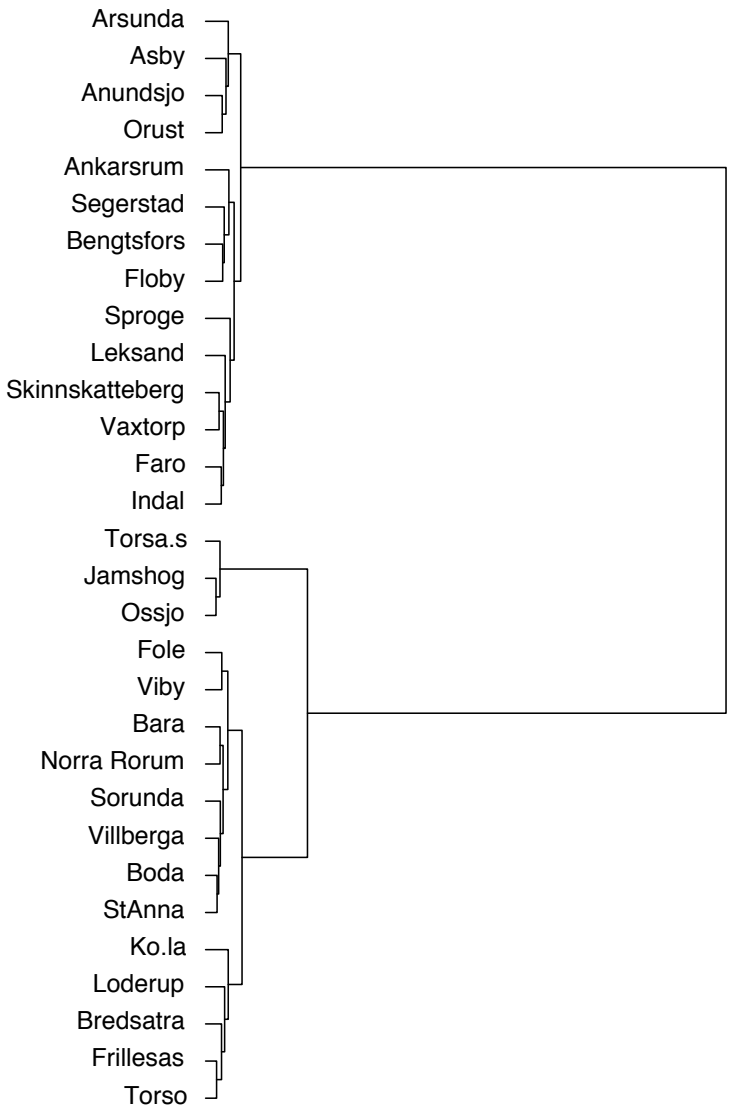


Figure 4.4: Dendrogram with cosine measure and trigram features, 5 normalizations

Consensus Trees

Consensus trees combine the results of cluster dendrograms, retaining only clusters that occur in the majority of dendrograms. When dendrograms have high agreement, the resulting consensus tree will retain most of the detail. When dendrograms have low agreement, the resulting consensus tree will be fairly flat. This avoids the dendrograms' problem of instability, where small changes in distances cause large re-arrangements in the tree. Only dendrograms whose input distances were at least 95% significant were used. That is, a measure/feature set combination had to be non-bold in tables 4.3 to 4.6 to be included. The consensus tree for full-site comparisons and 5 rounds of normalization is not given because there is only one dendrogram that qualifies.

It's worthwhile to note that more dendrograms were used to build the consensus tree of figure 4.7 than were used in figures 4.5 and 4.6. Despite this, figure 4.7 retains much more detail, indicating that its constituent dendrograms, based on 5 rounds of normalization, agree more than those with only 1 round of normalization.

The consensus trees are also grouped into clusters, which are then mapped in figures 4.8 – 4.10. The outline map of Sweden was provided by Therese Leinonen and is the same as those in Leinonen (2008). The L04 package from the University of Groningen was used to map the consensus trees onto the map of Sweden; the multi-dimensional scaling maps and composite cluster maps also used L04 (Kleiweg et al. 2004).

Analysis

The cluster dendrograms are dangerous to interpret too closely on their own; the instability of a single dendrogram means that small clusters cannot be analyzed reliably. For example, in figure 4.2, a two-way split between the sites on the top and bottom of the page is obvious, and another in the top cluster is easy to argue for, but outliers like Anundsjö and Årsunda are likely to shift from group to group in other dendrograms.

It is safer to analyze the consensus trees; the smoothing effect of taking the majority rule of each cluster will show where the optimal cutoff for splitting clusters is removing spurious detail. The

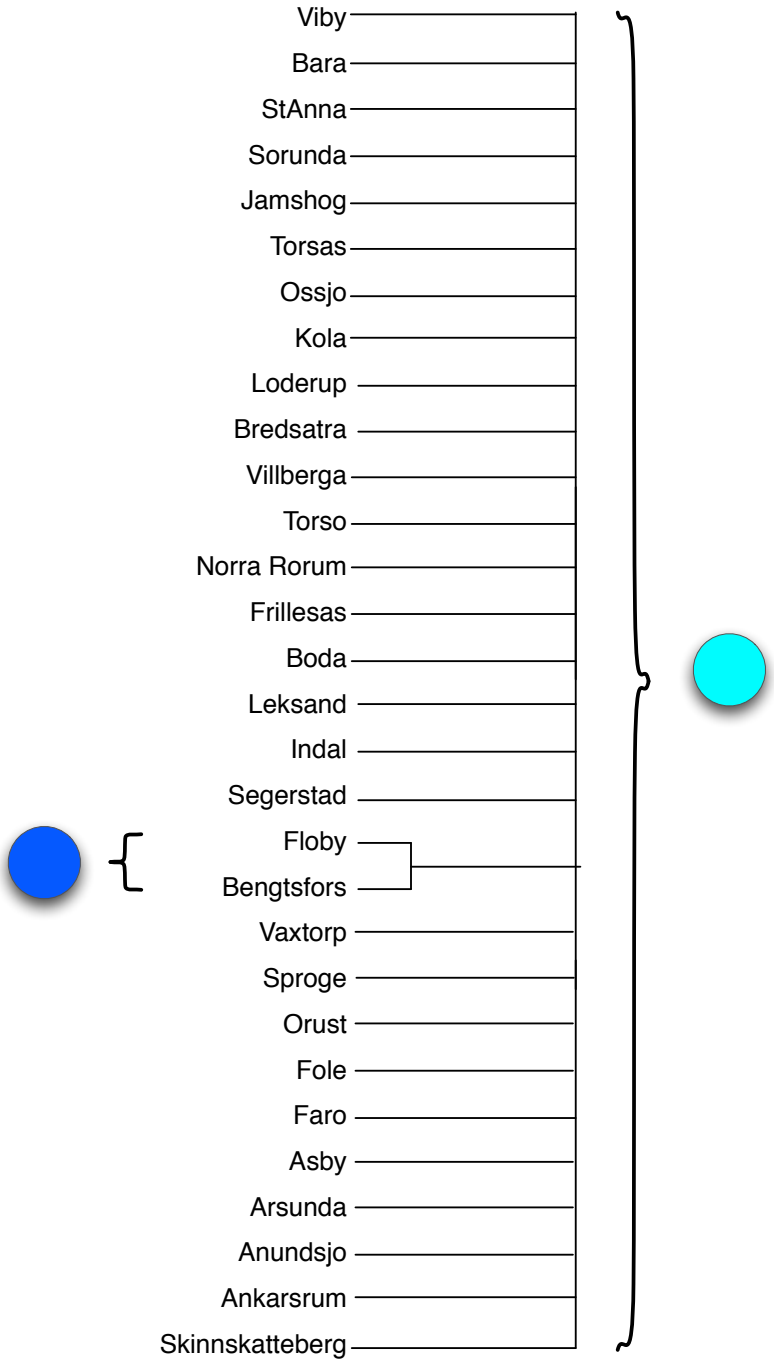


Figure 4.5: Consensus Tree for 1000-samples and 1 normalization

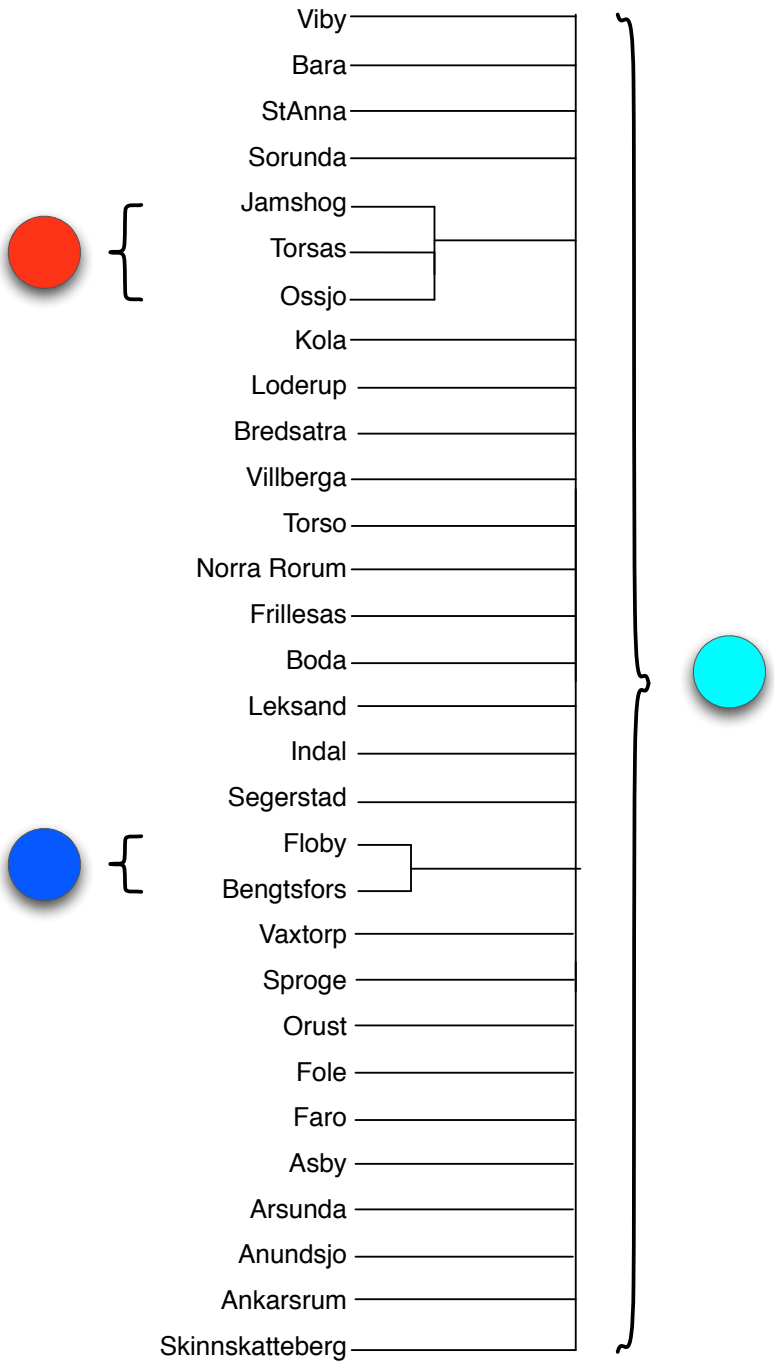


Figure 4.6: Consensus Tree for full site comparison and 1 normalization

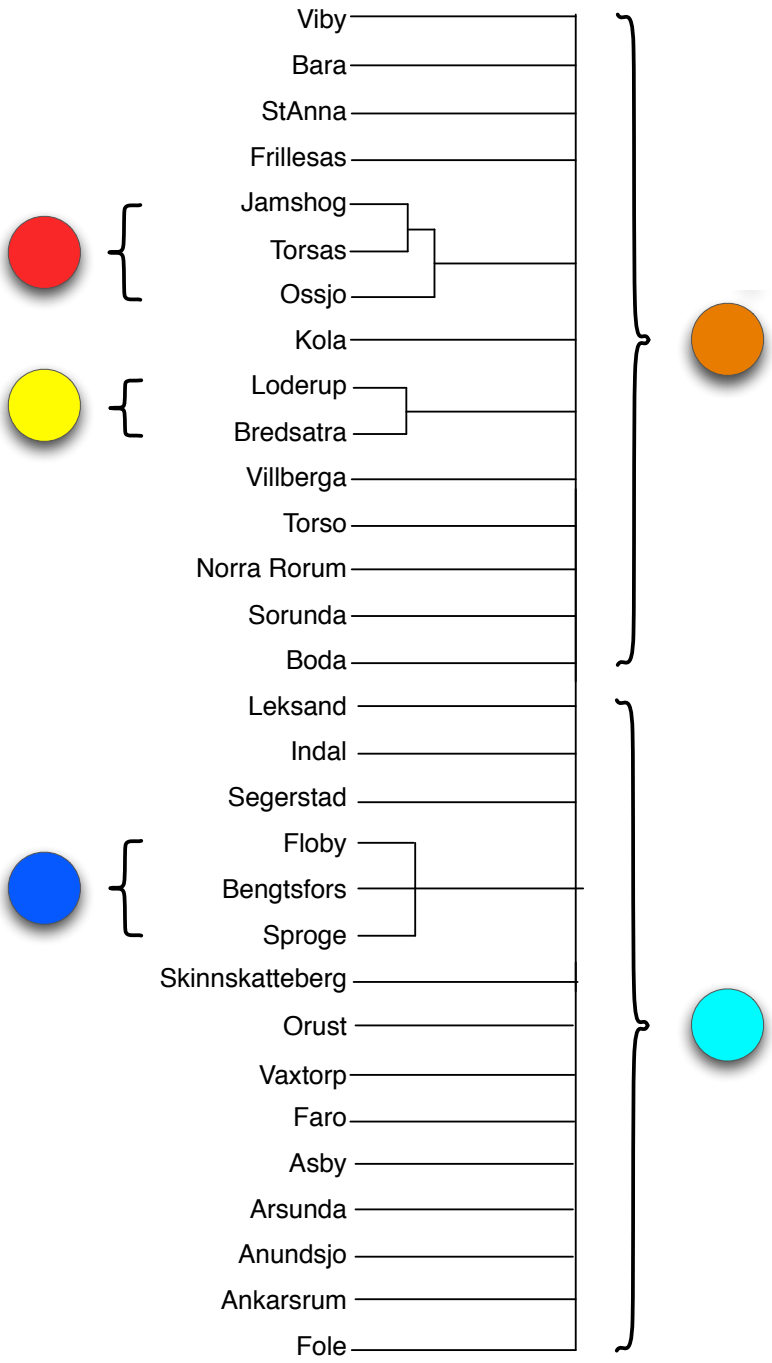


Figure 4.7: Consensus Tree for 1000-samples and 5 normalizations

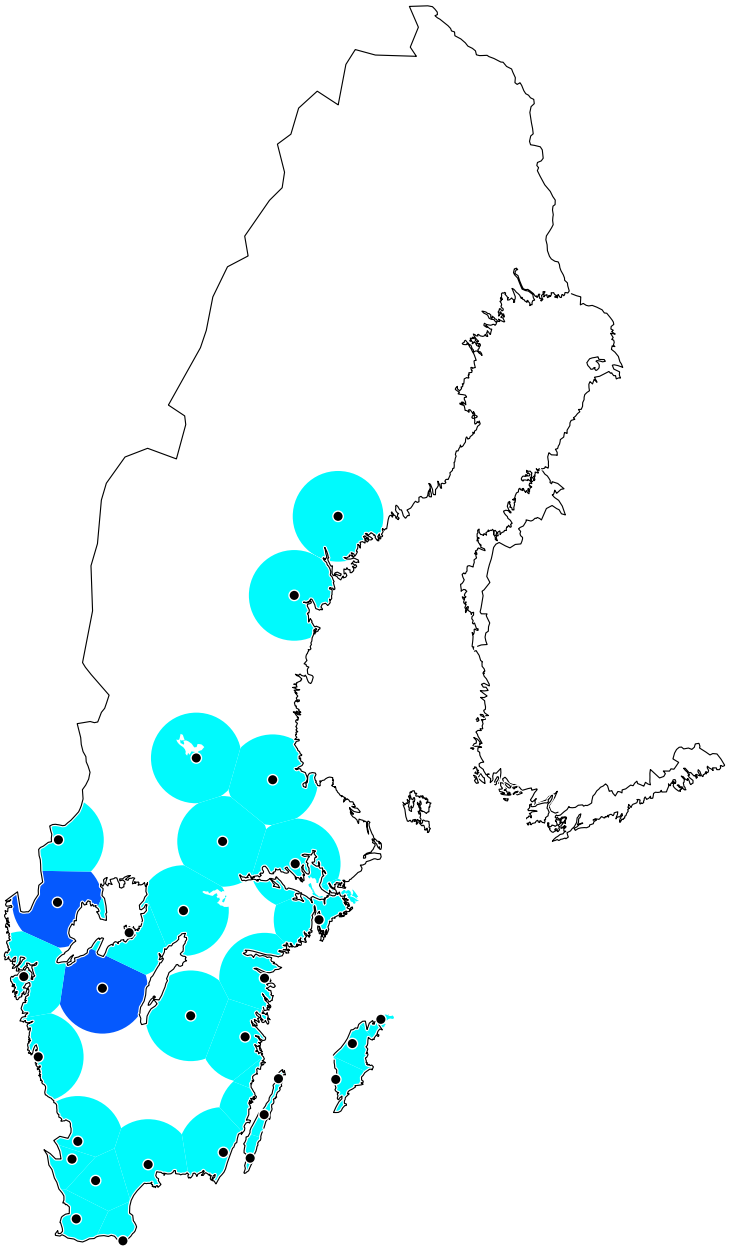


Figure 4.8: Consensus Tree for 1000-samples and 1 normalization, Mapped

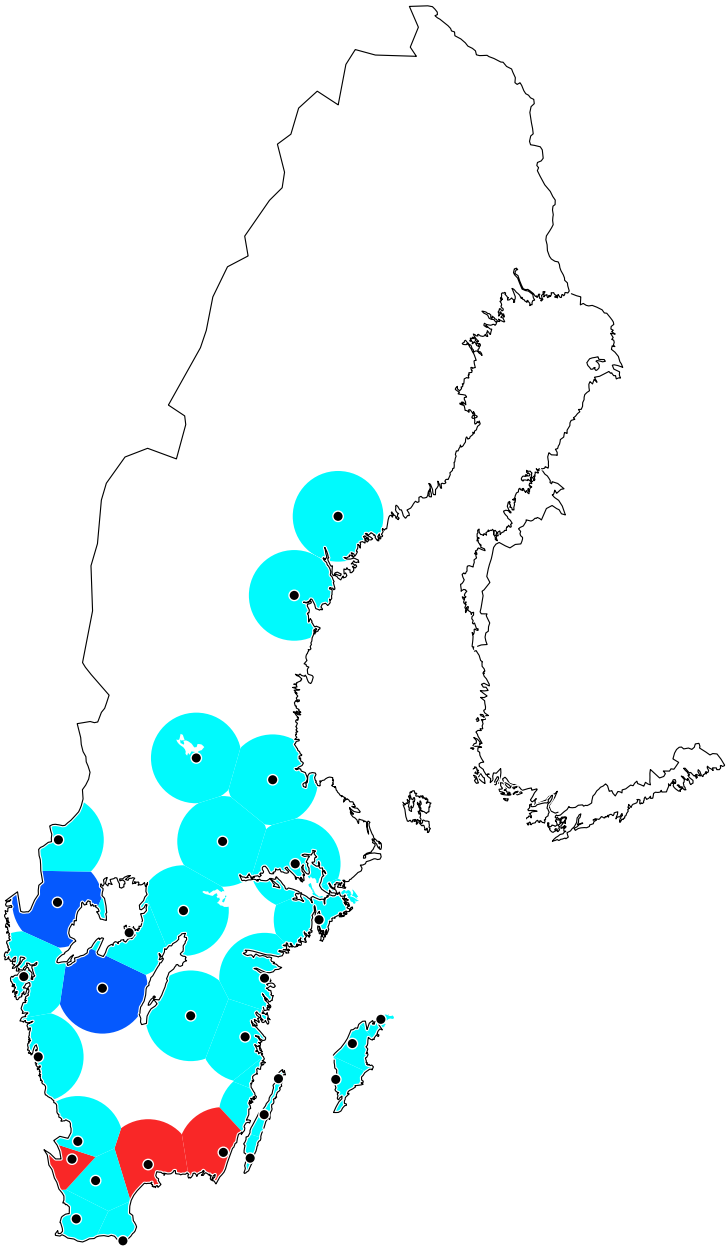


Figure 4.9: Consensus Tree for full site comparison and 1 normalization, Mapped

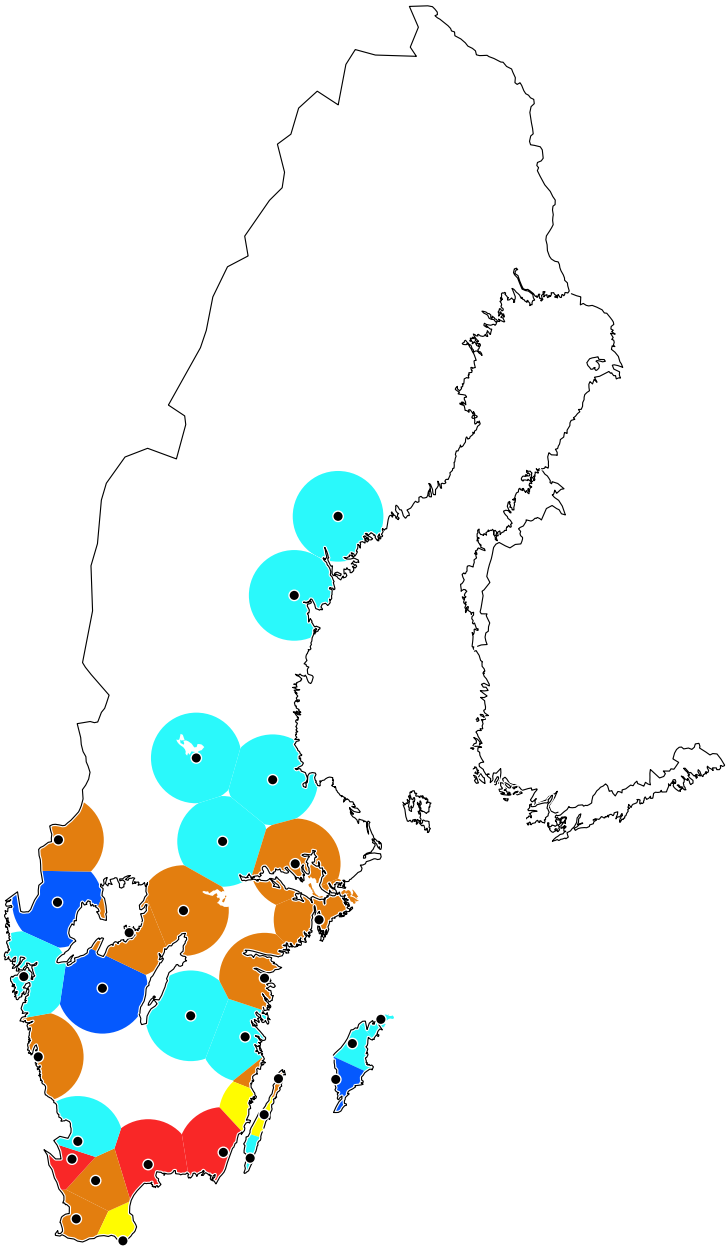


Figure 4.10: Consensus Tree for 1000-samples and 5 normalizations, Mapped

- Floby
- Bengtsfors
- Sproge (for 1000-sample, 5-normalization)

Figure 4.11: Blue Cluster

- Jämshög
- Torsås
- Össjö

Figure 4.12: Red Cluster

three consensus trees in figures 4.5 – 4.7 vary in amount of detail but the trees with more clusters do not contradict the clusters of the flatter trees.

For 1000-sentence samples and 1 round of normalization, there is one cluster: Floby and Bengtsfors. Full-site comparison finds another cluster: Jämshog, Össjö and Torsås. Finally, 1000-sentence samples and 5 rounds of normalization finds another cluster consisting of Löderup and Bredsätra. It also finds a large two-way split between the sites and adds Sproge to the first cluster with Floby and Bengtsfors. To aid further analysis, the clusters are assigned colors, which are detailed in figures 4.11 – 4.15.

When these clusters are mapped onto the geography of Sweden, some patterns are visible. Since figure 4.7 is strictly more complex than the preceding two, it is used as the basis for this analysis–see figure 4.10. The large two-way split is between the orange and cyan clusters. The orange cluster, which includes red and yellow clusters, forms two horizontal bands across Sweden. The centers of the orange cluster appear to be Stockholm and Malmö. Meanwhile, the red and yellow clusters

- Bredsätra
- Löderup

Figure 4.13: Yellow Cluster

- Leksand
- Indal
- Segerstad
- Floby
- Bengtsfors
- Sproge
- Skinnskatteberg
- Orust
- Våxtorp
- Fårö
- Asby
- Årsunda
- Anundsjö
- Ankarsrum
- Fole

Figure 4.14: Cyan Cluster

- Viby
- Bara
- S:t Anna
- Frillesås
- Jämshog
- Torsås
- Össjö
- Köla
- Löderup
- Bredsätra
- Villberga
- Torsö
- Norra Rörum
- Sorunda
- Böda

Figure 4.15: Orange Cluster

form a boundary along the northern border of Skåne and Blekinge counties.

Meanwhile, the cyan cluster, which includes the blue cluster, seems to represent the countryside of Sweden. On the other hand, because the blue cluster is near Göteborg, it might be better characterized simply as "non-Stockholm". This matches the traditional dialect regions of Sweden, with the exception of of the city/country divide, and the fact that this hard clustering simplifies the dialect boundaries, which are traditionally believed to be gradient. Also, the island Gotland is not put in a separate cluster as predicted by traditional boundaries. For discussion, see section 5.1.

Composite Cluster Maps

Composite cluster maps use an underlying technique similar to consensus trees–cluster dendrograms, but they combine and present the information in a very different way. They, too, provide a stabler view of the groups that sites form when clustered. This view, however, emphasizes the boundaries between sites. The result looks much more like the traditional isogloss boundaries of dialectology.

The three composite cluster maps in figures 4.16 – 4.18 are the composite of the same dendrograms used as input for the consensus trees: all-significant parameter settings, divided by type of normalization (sentence-length only or ratio added as well).

All three composite clusters maps provide a picture similar to the consensus tree map 4.10 of the previous section. The north-to-south gradient is supported by the weak horizontal boundaries present up and down Sweden.

Of these boundaries, the one between Skåne and the rest of Sweden is the strongest. Due to the lack of interview sites in the middle of south Sweden, the boundary is drawn further north than it traditionally appears, but this is an effect of the software that produced the figure. Notice that there is also a boundary between the red cluster, comprised of Jämshog, Torsås, and Össjö, and the other sites, especially visible in figures 4.16 and 4.18. Their presence along the northern border of Skåne is one reason why its boundary with the rest of Sweden is so strong.

Compared to the consensus tree maps, the composite cluster maps cannot support the

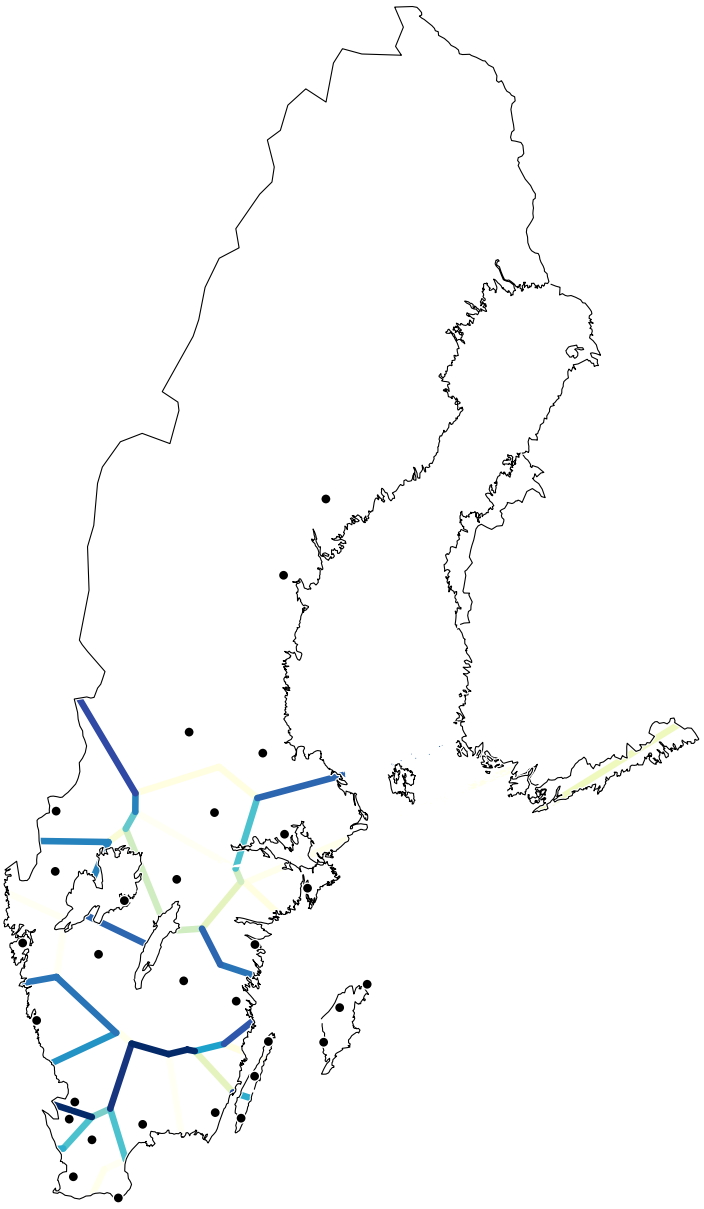


Figure 4.16: Composite Cluster Map for 1000-sample, 1 normalization

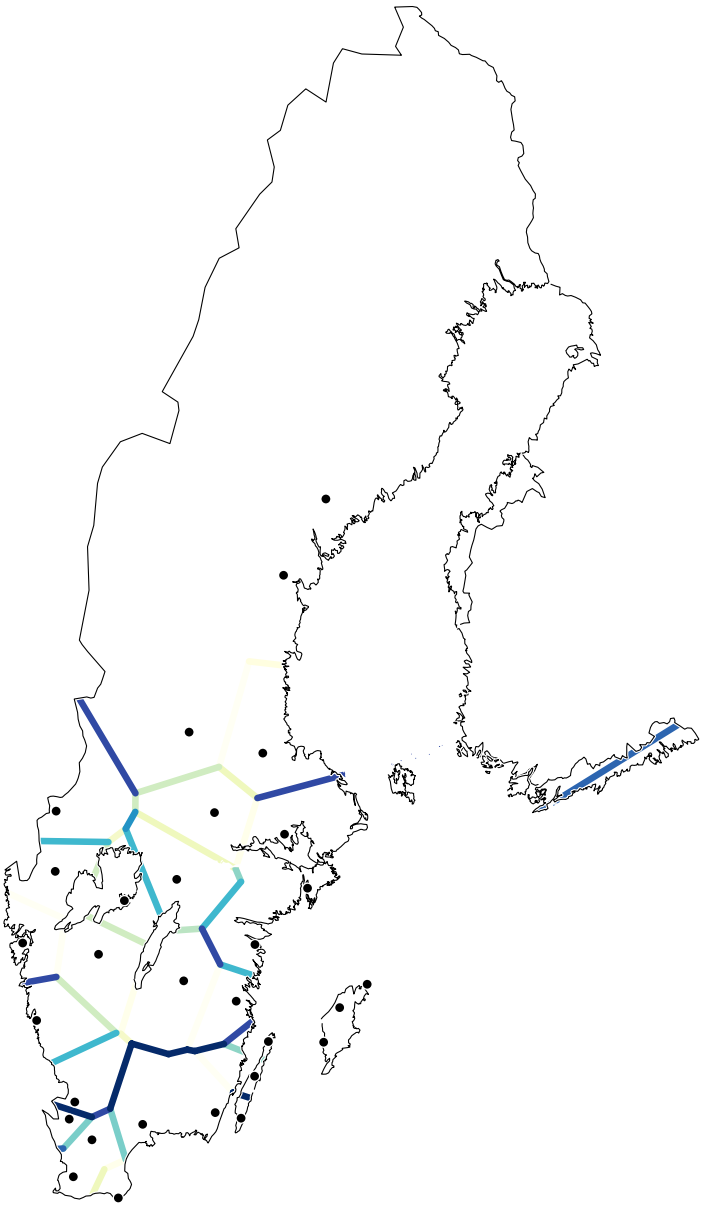


Figure 4.17: Composite Cluster Map for complete sites, 1 normalization

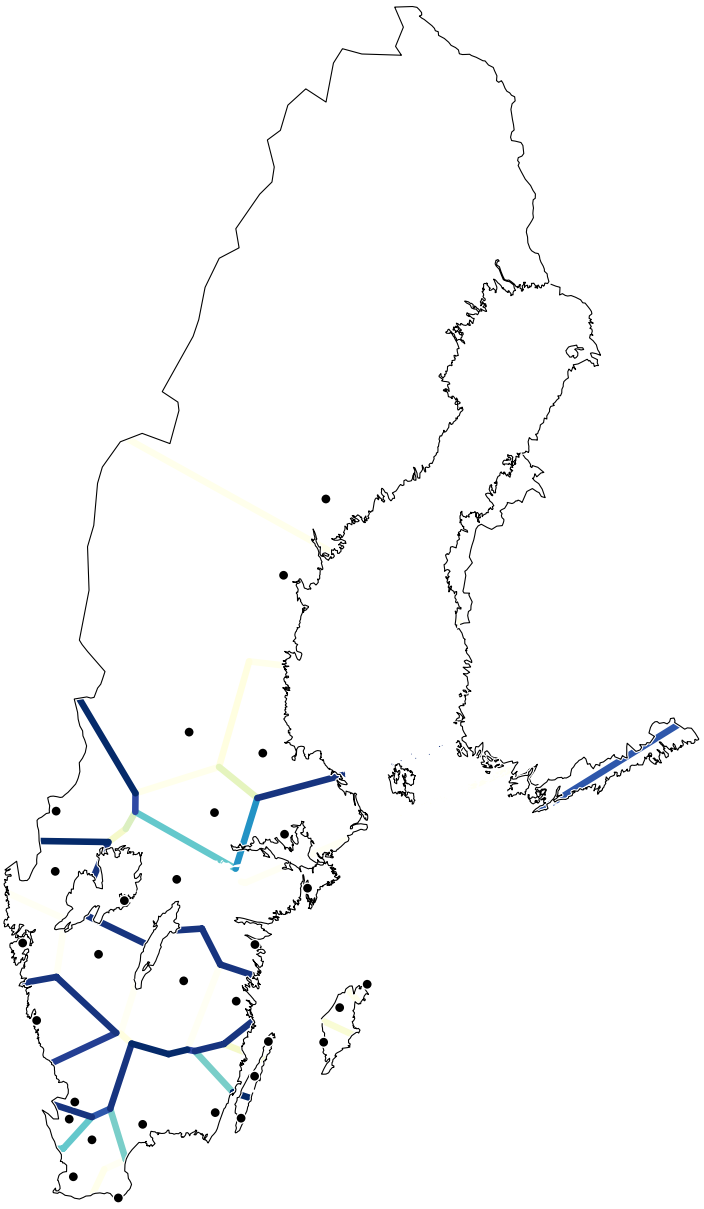


Figure 4.18: Composite Cluster Map for complete sites, 5 normalizations

city/country distinction because there is no way to identify distant areas by their color. On the other hand, it is possible to detect the relative strength of a boundary. To combine these two features, multi-dimensional scaling is needed.

4.5 Multi-Dimensional Scaling

Multi-dimensional scaling (MDS) plays a similar role to clusters, condensing the high-dimensional information into an form that is easier to understand. It differs, however, in producing gradient numbers, not binary trees: cluster dendrograms put each site into one and only one cluster, whereas MDS puts each site into 3D space; the clusters are only implicit in the positions. This also means that MDS maps are more stable than dendrograms.

Kruskal's (1964)b MDS works by positioning the dissimilarities in high-dimensional space, then converting them to true distances in some lower dimensional space—in this case, three dimensions. It distorts the dissimilarities equally and by the minimum amount necessary. Kruskal calls the measure of distortion Stress. Once the sites are points in 3D space, each sites' x, y, and z co-ordinates can be mapped onto the colors red, green, and blue, then drawn on a map of Sweden.

It must be noted that the maps vary in color because of the way that MDS positions the sites in 3D space, based on the distances between them. Kruskal's method guarantees that its results are comparable for equivalent inputs, but this may not always be obvious because the color equivalence may be difficult to decipher. Equivalent MDS maps may be rotated with respect to each other in 3D space, and this rotation is visible in the color selection: if two sites are both blue in one map and in another map are both orange, then they have the same relation to each other.

The maps shown here in figures 4.19 – 4.21 are based on the same parameter settings as the dendrograms in figures 4.1 – 4.3. The first two are Jensen-Shannon divergence measured over trigrams, with 1 and 5 rounds of normalization, respectively. The third is R^2 measured over phrase-structure rules with 1 round of normalization.

Despite the differences between MDS and the preceding methods, the similar results are evident; the maps (figures 4.19 – 4.21) all show the same patterns as the other methods. That is, there

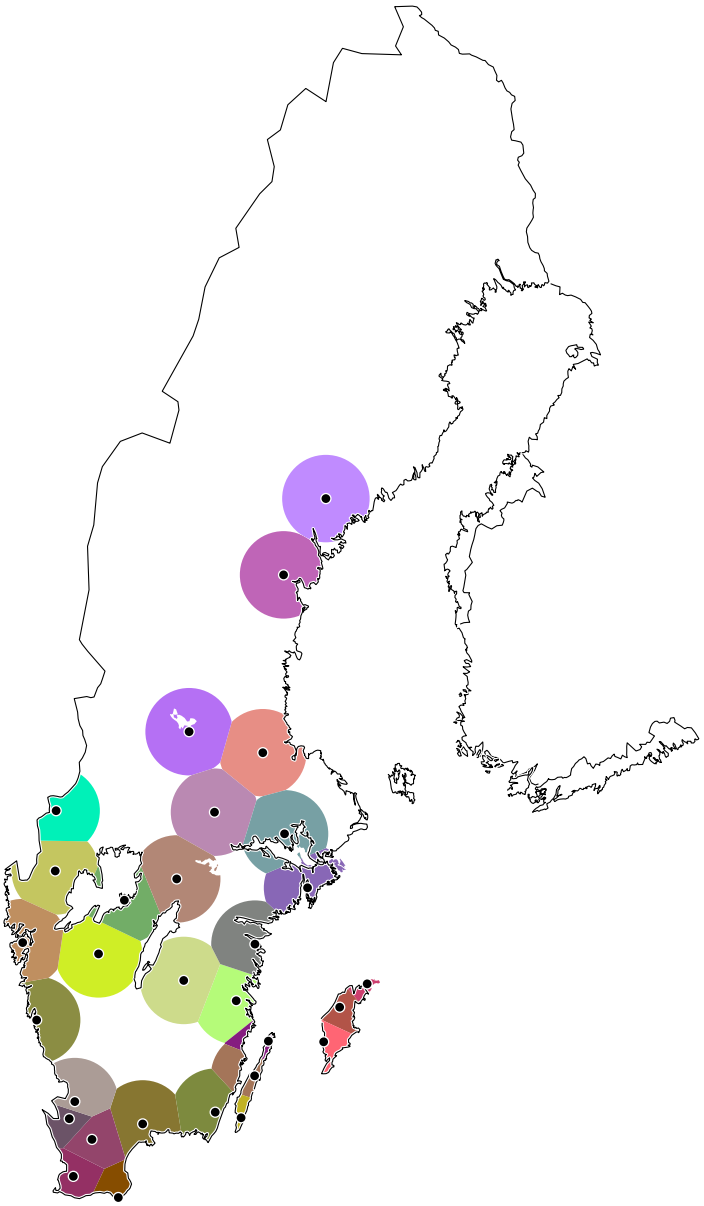


Figure 4.19: Jensen-Shannon measure with trigram features, 1000-sentence sampling and 1 round of normalization

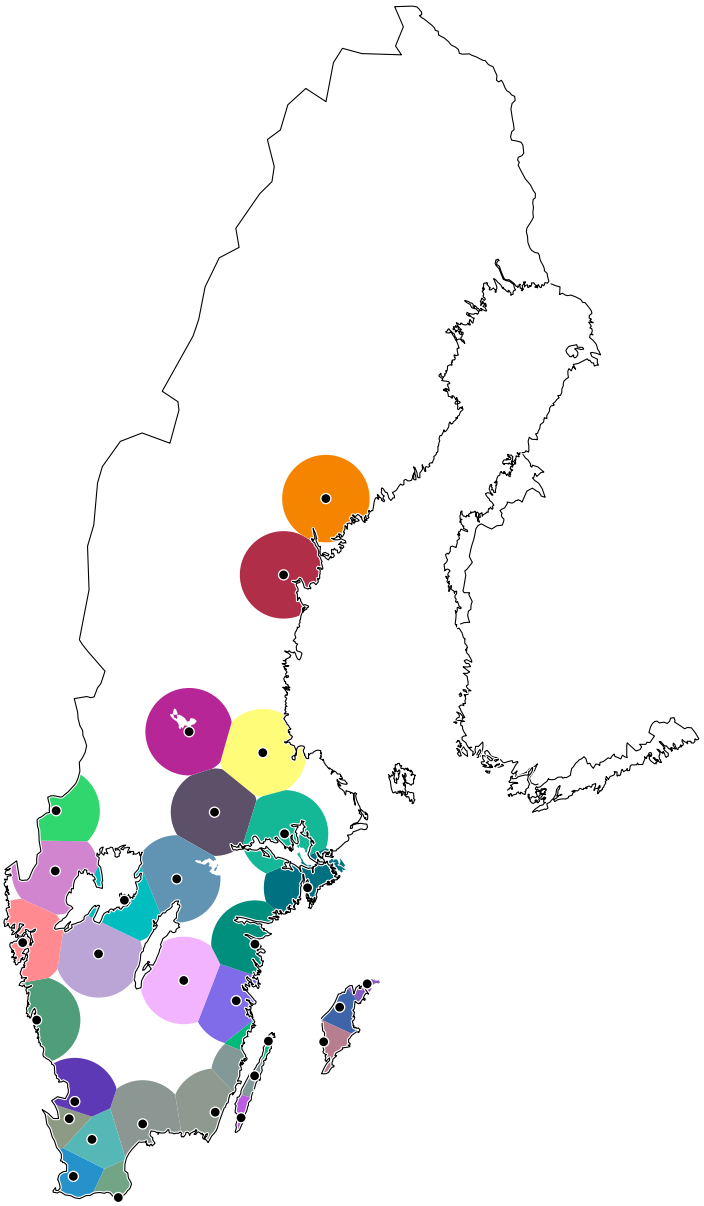


Figure 4.20: Jensen-Shannon measure with trigram features, 1000-sentence sampling and 5 rounds of normalization

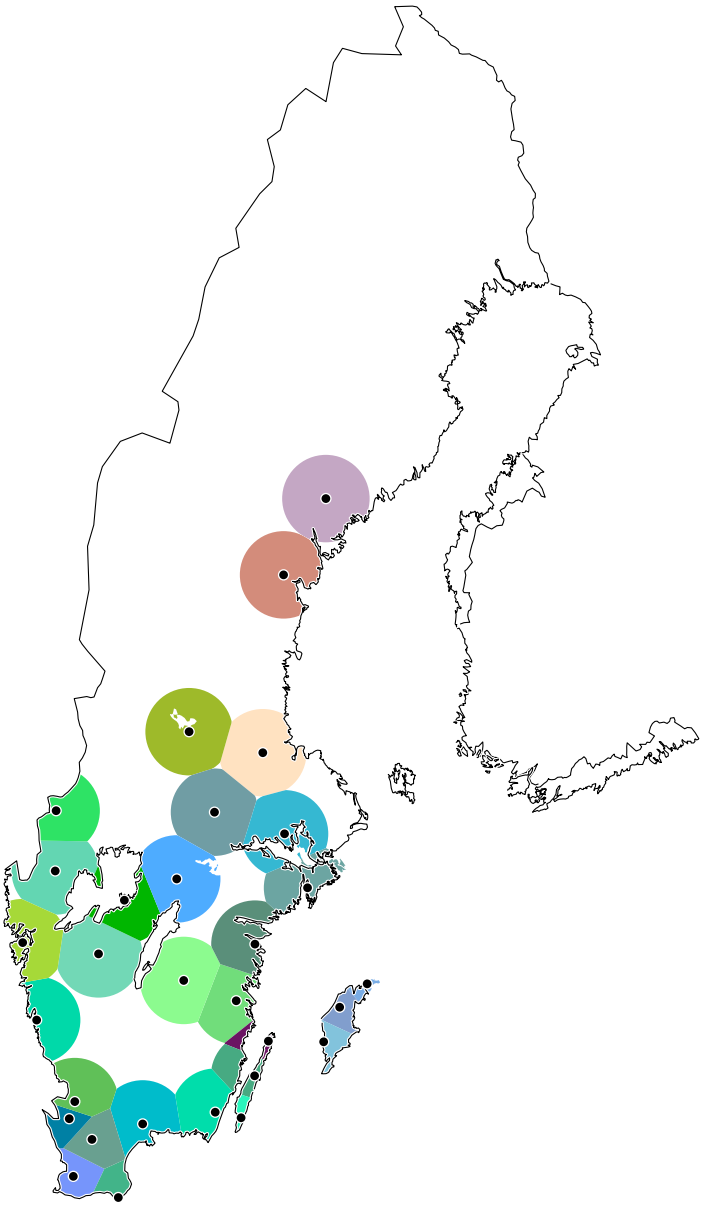


Figure 4.21: R^2 measure with phrase-structure-rule features, full-site comparison and 1 round of normalization

is a general north-to-south gradience, especially easy to see in map 4.19. There is a strong southern cluster, visible in all of the diagrams. And there is a general two-way distinction between city and country.

The main contribution that the MDS maps make is that the north-to-south gradient is more obviously gradient. In other words, it is easier to see the gradation from north to south. For example, in figure 4.21, looking from the north to south, the colors change quickly close to Stockholm, then fade to green further south, then transition back to blues and purples further south, in Skåne.

The Stockholm and Malmö areas, which are in the same cluster in the consensus tree maps, are here seen to be similar without being identical. For example, in figure 4.20, the Stockholm area is a shade of blue-green while the Malmö area is a shade of blue-grey. Also in figure 4.20, Skåne and Blekinge are grey: clearly similar but not identical to Malmö.

4.6 Features

Ranked features answer the question of agreement with dialectology more precisely than the previous two methods. Features are ranked by their normalized weight, which tells how much weight the distance measure will give it. This can reveal aggregate differences that may only be noticeable when counting a large amount of data. Conversely, with different normalization settings, feature ranking can also point out rare features that only occur in one site. The first kind of features are unlikely to be noticed by linguists without the aid of computers, whereas the second kind are the rare features that are easy for linguists to notice.

Both kinds of normalizations are shown below; in the first set of rankings, the normalizations for sentence size are applied, as described in section 3.2, whereas the second set of ranking is normalized for relative overuse, based on Wiersma's (2009) normalization described in

4.7 wiersma-normalization

. The overuse normalization shows which features are used relatively more when comparing two sites. It normalizes for feature frequency between two sites.

Without the overuse normalization, the top-ranked features will tend to be the most common ones, those found in almost every sentence in the interview. These common features tend to highlight gradient differences: differences in quantity but not in quality. In contrast, the overuse normalization allows us to see which features happen only a few times in one side of the comparison and not at all in the other. This is closer to a traditional linguistic analysis.

In addition, only features that appear in both groups being compared were ranked; although features that only appear in one or the other can be interesting, they tend to be noisy in features extracted from automatically annotated corpora. It is not possible to tell which unique features are interesting and which are noise, especially when using the overuse normalization, which makes rarely occurring features rank similarly to common ones.

These results compare clusters from the consensus trees based on 1 round of normalization (figures 4.5 and 4.6) as well as the consensus tree based on 5 rounds of normalization and a 1000-sentence sample (figure 4.7). The consensus tree for 5 rounds of normalization and full-site comparisons only had one tree for input and was not usable. Given these three consensus trees, the groups in table 4.20 are the relevant ones for analysis.

There are four clusters, three small and one large which contains the remainder of the sites. They are listed in table 4.20. Cluster A, containing Floby and Bengtsfors, appears in all three consensus trees. Its features are colored blue in the following figures. Cluster B, containing Jamshog, Torsas and Ossjo, appears in the second two trees. It features are colored red. Cluster C, containing Loderup and Bredsatra, appears only in the third tree. Its features are colored yellow. The remainder of the sites are in Cluster D; the third consensus tree differs from the first two in splitting the remainder into two groups, but this division is ignored here to reduce the number of comparisons. Between large groups of sites, such comparisons are unlikely to be informative anyway.

A	(Blue) Floby, Bengtsfors
B	(Red) Jämshog, Össjö, Torsås
C	(Yellow) Löderup, Bredsätra
D	(Cyan) Segerstad, Köla, S:t Anna, Sorunda, Norra Rorum, Villberga, Torso, Boda, Frillesås, Indal, Leksand, Anundsjö, Årsunda, Asby, Orust, Våxtorp, Fole, Sproge, Fårö, Ankarsrum, Skinnskatteberg

Table 4.20: Clusters discussed

For each pair of clusters, I rank and analyze the input features by comparing feature differences. The features presented here are the ten highest ranked features for a particular comparison. Although each feature set has ten features ranked here, they are better thought of as two sets of five features differences. The top five positive features are shown as are the top five negative features, scaled such that the most important feature has the value 1.0.

This has two advantages. It splits the features so that both the positive and negative evidence are always visible; otherwise, in some cases, if one side is strong enough, the other would be pushed out of the top ten. However, it still allows the relative weight of evidence to be estimated. For example, if some cluster has some idiosyncratic features, most of the features will be positive, meaning that features typical of that cluster contribute most to the distance between it and other clusters. The two-part feature will show this: the five positive features will have much higher values than the five negative features.

The first subsection, 4.7, shows all comparisons between clusters for a single parameter setting: trigram features, 1000-sentence sampling and sentence-size normalization only. Besides unigrams, these are the parameters that give the highest correlation with travel distance for 1000-sentence sampling. In the next subsection, 4.7, the overuse normalization is added, keeping other parameter settings the same. The third subsection, 4.7, a single comparison between cluster A and cluster B is given for all feature sets. In the final subsection, 4.7, the high-ranked phrase-structure rules are given. The parts of speech for the features are given in table 4.21. The non-terminals are given in table 4.22.

POS	Part of speech	POS	Part of speech
++	coordinating conjunction	MV	verb “måste” (must)
AB	adverb	NN	noun
AJ	adjective	PN	proper name
AN	adjectival noun	PO	pronoun
AV	verb “vara” (be)	PR	preposition
BV	verb “bli(va)” (become)	QV	verb “kunna” (can)
EN	indefinite article	RO	numeral
FV	verb “få” (get)	SP	present participle
GV	verb “göra” (do)	SV	verb “skola” (shall)
HV	verb “hava” (have)	UK	subordinating conjunction
I?	question mark	VN	verbal noun
ID	idiom	VV	other verb
IM	infinitive marker	WV	verb “vilja” (want)
IP	period	XX	Unclassifiable
MN	meta-noun	YY	Interjection

Table 4.21: List of parts of speech

NT	Non-terminal
+F	Coordination at main clause level
AA	other adverbial
AP	adjective phrase
AVP	adverb phrase
CAP	Coordinated adjective phrase
CNP	Coordinated noun phrase
CONJP	Other coordinated phrase
CS	Coordinated S
ET	Other nominal post-modifier
MS	Macrosyntagm
NAC	Not a constituent
NP	Noun phrase
OA	Object adverbial
PP	Prepositional phrase
RA	Place adverbial
S	Sentence
SS	Other subject
TA	Time adverbial

Table 4.22: List of non-terminal labels

Trigram Features

The analysis will start with trigram features without the overuse normalization, since trigrams have the highest rate of significance of the non-combined feature sets. Clusters are compared with other, single clusters in order to present both the features that are more common in a cluster as well as those that are less common; negative comparisons against combined clusters are less informative because the most common features over the whole set overpower the interesting features from the single cluster being compared. In addition, the combined feature set is not presented because the mixed feature types make it difficult to interpret.

As mentioned above, the top-ranked trigrams are common, typical of the core of the sentence. The trigrams typical of cluster A are formed most often, for example, from a trigram like *"och det är"* "and it's", followed by an adverb such as *"väl"* "well". Figures 4.22 – 4.24 generalize this observation by the high rankings of the POS trigrams PO-AV-AB (pronoun-copula-adverb, figure 4.22), ++-PO-AV (conjunction-pronoun-copula, figure 4.22), and PO-VV-AB (pronoun-verb-adverb, figure 4.24) The same is true of the other clusters for the most part. Unfortunately, this makes it hard to say interesting things about the difference in feature distribution. It does appear that clusters B and C use adverbs and of conjunctions that differ from the other clusters; for example ++-AB-AV in in figure 4.23. The comparison between cluster A and cluster B highlights the trigram AB-AB-AB (figure 4.22) as important, but more interesting are the ++-AB-AV (conjunction-copula-adverb) and AB-AV-AB (adverb-copula-adverb) trigrams in the bottom halves of figures 4.22 and 4.23. These trigrams derive from sequences like *"och så är"* ("and is so") and *"inte är ju"* ("is not now").

Trigrams with Overuse Normalization

Given this lack of information, there are two dimensions along which the comparisons can be altered: normalization and feature set. Starting with normalization, let us add the overuse normalization technique. Differences appear immediately. First, the balance of feature weight obviously differs here. For example, in the comparison between cluster A and cluster B (figure 4.28), the features of cluster A are more important in distinguishing the two than the features of cluster B. The comparison between cluster A and cluster D (figure 4.30) is so lop-sided that cluster

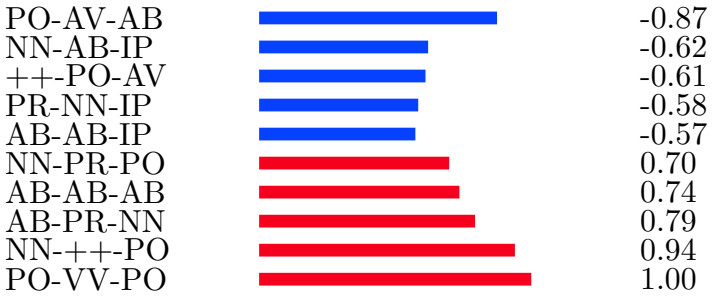


Figure 4.22: cluster A ⇔ cluster B, trigram features

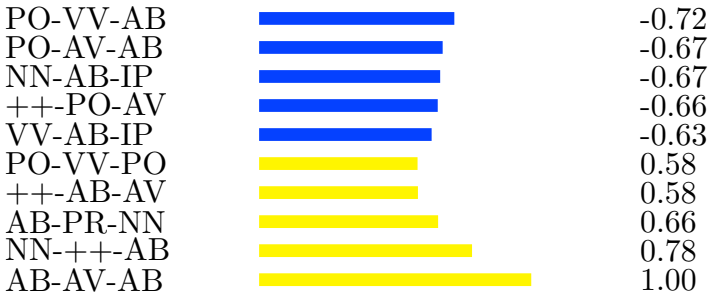


Figure 4.23: cluster A ⇔ cluster C, trigram features

Figure 4.24: cluster A ⇔ cluster D, trigram features

Figure 4.25: cluster B ⇔ cluster C, trigram features

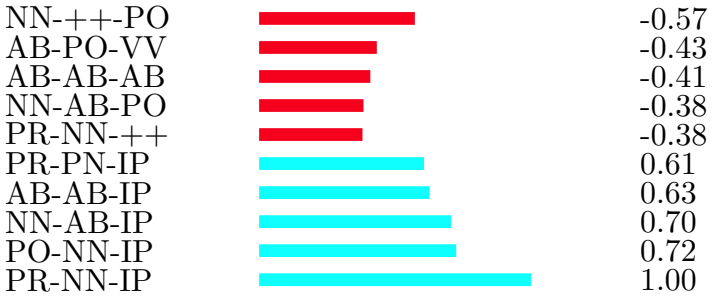


Figure 4.26: cluster B ⇔ cluster D, trigram features

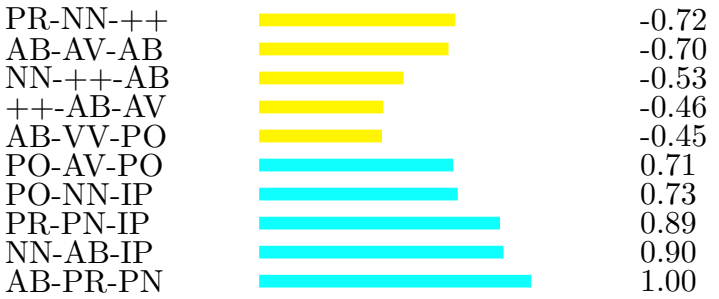


Figure 4.27: cluster C ⇔ cluster D, trigram features

D contributes no features at all. This occurs when none of the features the two clusters share are overused in cluster D.

With the overuse normalization, cluster A has two interesting patterns. First, the trigrams it overuses are filled with indefinite articles (EN) and prepositions (PR). Examples include VV-EN-AB (verb-indefinite-adverb), PR-EN-AB (preposition-indefinite-adverb) and PR-EN-VN (preposition-indefinite-verbal noun) in figure 4.28, as well as IM-PR-NN (infinitive marker-preposition-noun) and PR-ID-PR (preposition-idiom-preposition) in figure 4.30. These trigrams, such as PR-EN-AB, arise from sequences like *"om en inte"* "if one [does] not" or PR-EN-VN from *"ifrån en tävling"* "from a competition".

Second, the trigrams it underuses mostly end with pronouns: 4 of 5 trigrams in the comparison with cluster B (figure 4.28) and 4 or 5 in the comparison with cluster C (figure 4.29). Even in the comparison with cluster D (figure 4.30), 4 of 5 of the "least overused" trigrams end with pronouns. (The low values in the bottom half of the comparison with cluster D are not underused by cluster A, because cluster D has no unique features here. Instead they are the "least overused" by cluster A.)

Cluster B shows one interesting pattern: overuse of sköla (shall), including an interesting trigram SV-QV-AB (shall verb-can verb-adverb) in figure 4.32. Although this could be a mistake on the part of the tagger, the different forms of this verb are limited, so this is unlikely: identifying them is not hard. An example utterance with this pattern is *"För det var . . . "* "For it was that one . . . ". Here the construction *"skulla kunna"* appears to be a double modal, similar to the English double modal "should can".

Cluster C doesn't gain any interesting patterns with overuse normalization in figures 4.29, 4.31, and 4.33, except for a surprising variety in the verbs: göra (do), hava (have), kunna (can), sköla (shall), vara (be) and vilja (want). Many uses of adverbs show up as well. It is not clear what either of these patterns mean linguistically, however.

Cluster D gives no information whatsoever when the overuse normalization is added, simply because it has no informative features. This is expected, given its nature as a combination of many sites. The tradeoff of more informative features for the smaller clusters is worthwhile.

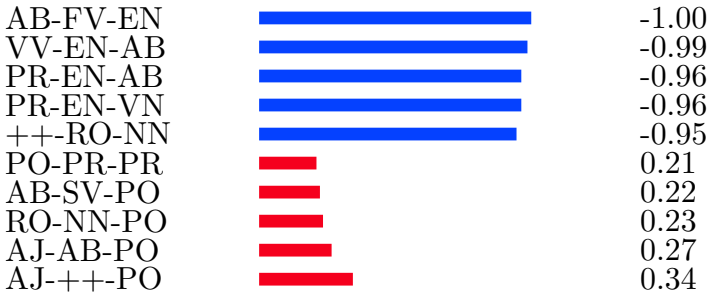


Figure 4.28: cluster A ⇔ cluster B, trigram features with overuse normalization

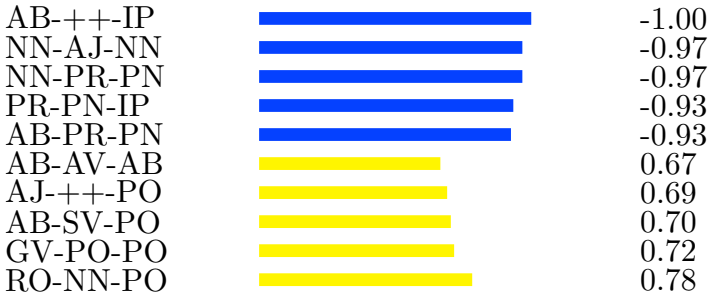


Figure 4.29: cluster A ⇔ cluster C, trigram features with overuse normalization

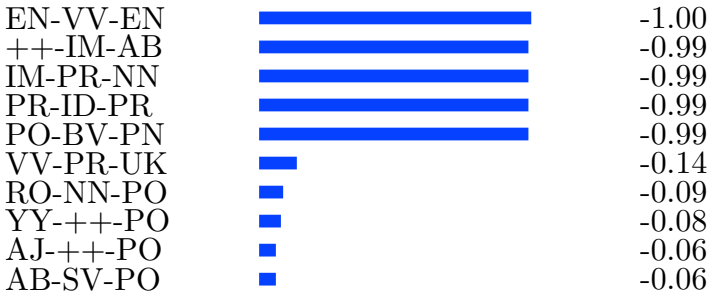


Figure 4.30: cluster A ⇔ cluster D, trigram features with overuse normalization

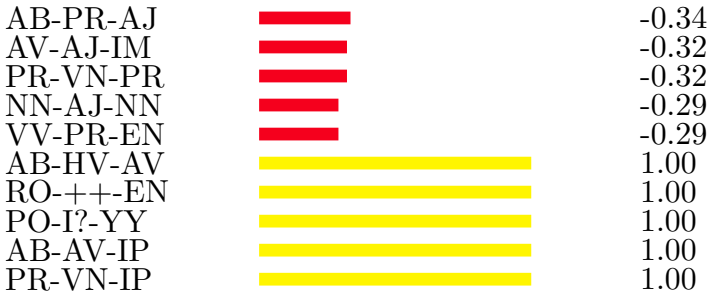


Figure 4.31: cluster B ⇔ cluster C, trigram features with overuse normalization

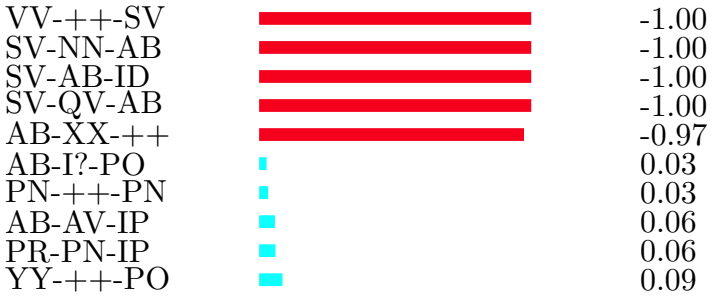


Figure 4.32: cluster B ⇔ cluster D, trigram features with overuse normalization

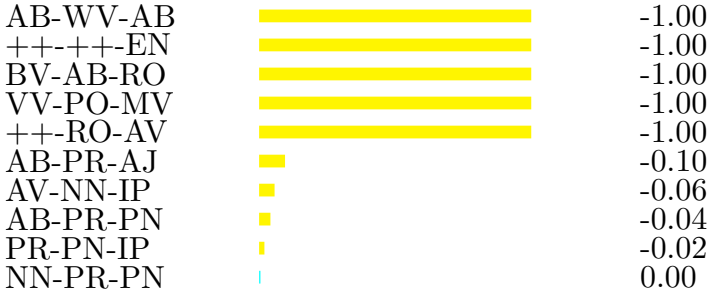


Figure 4.33: cluster C ⇔ cluster D, trigram features with overuse normalization

Variation Across Feature Sets

Moving to other feature sets with overuse normalization, leaf-ancestor paths and leaf-head paths, figures 4.34 and 4.40, give additional information about cluster A that lead to the conclusion its defining characteristic is simple sentences, simpler at least than the other clusters. Specifically, cluster A's overused leaf-ancestor paths include few nested sentences (figure 4.34). This contrasts sharply with cluster B and cluster C in figures 4.37 – 4.39, which include many nested sentences. Cluster A does have complex paths, but they feature prepositional phrases. (Note: NAC stands for "not a constituent" and indicates that the parser could not decide what the correct constituent was at that point, or that there are crossing branches, which is less common.)

This characteristic of cluster A appears in the leaf-head paths as well (figure 4.40); cluster A's paths contain many [adjective]-noun-preposition sequences, but few verb-verb sequences that indicate nested phrases. Again, cluster B and cluster C (figures 4.37 – 4.39) have many of these sequences. Both clusters have a number of overused adverb features as well, similar to the trigram results. Note that comparison to cluster D is less interesting. Because it has fewer unique characteristics, when compared to it, clusters A, B and C show more generic characteristics. For example, all three clusters show that their sentences are generally more complex than the general sites in cluster D. This may be a sign that the normalizations are not fully working; cluster D is larger than A, which is larger than clusters B and C, so it seems that larger sets of sites are decided based on simpler features.

Phrase-structure rule features

Analysis of the phrase-structure-rule features is difficult because of all the noise. In figures 4.46 and 4.48, features like S→ ++-AB (conjunction-adverb) S→FV-PO-AB-VV (get verb-pronoun-adverb-verb) are hard to describe as anything but junk rules created by the parser. On the other hand, there are a lot of linguistically odd but reasonable rules like S→PO-AV-NP-IP (pronoun-copula-noun phrase-period) in figure 4.47. Although this is not a good linguistic decomposition, it is one that a statistical parser would create when copular sentences are common enough.

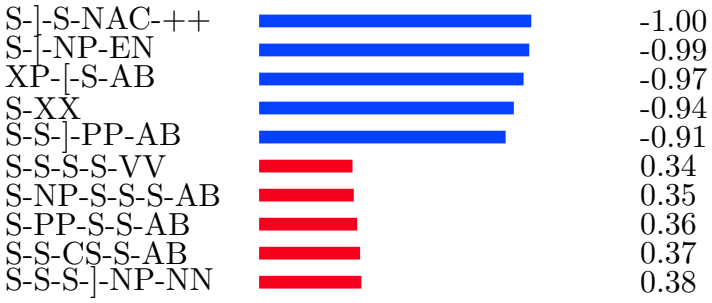


Figure 4.34: cluster A ⇔ cluster B, leaf-ancestor path features

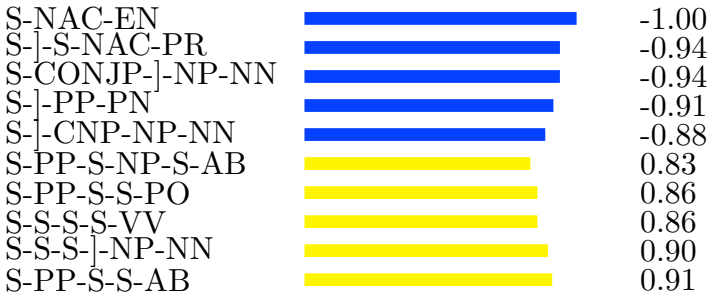


Figure 4.35: cluster A ⇔ cluster C, leaf-ancestor path features

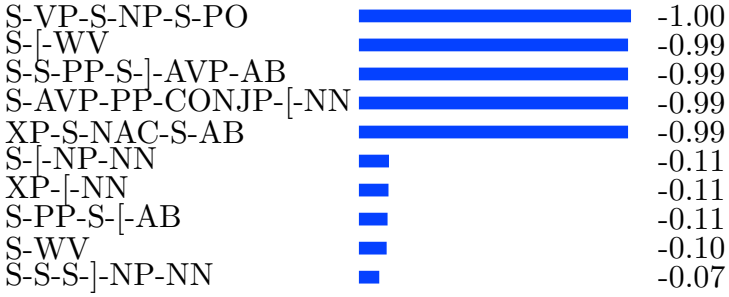


Figure 4.36: cluster A ⇔ cluster D, leaf-ancestor path features

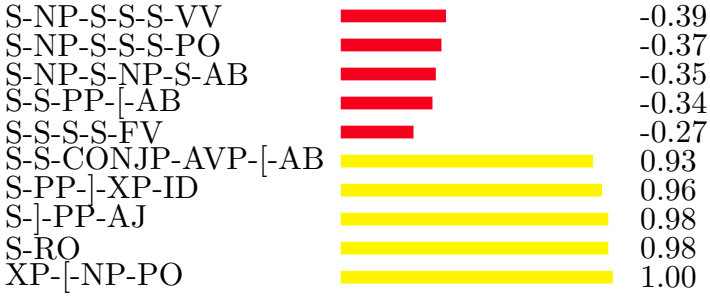


Figure 4.37: cluster B ⇔ cluster C, leaf-ancestor path features

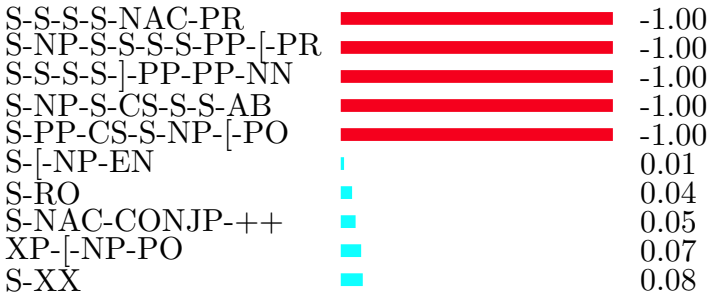


Figure 4.38: cluster B ⇔ cluster D, leaf-ancestor path features

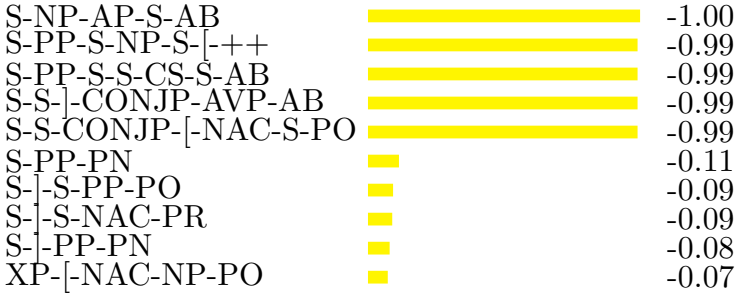


Figure 4.39: cluster C ⇔ cluster D, leaf-ancestor path features

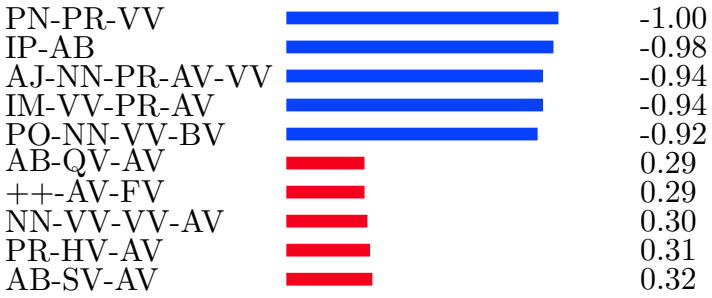


Figure 4.40: cluster A ⇔ cluster B, leaf-head features

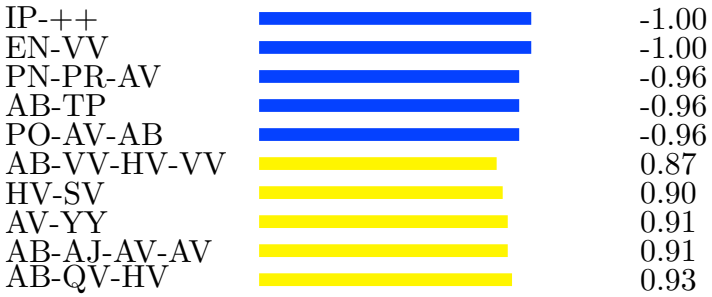


Figure 4.41: cluster A ⇔ cluster C, leaf-head features

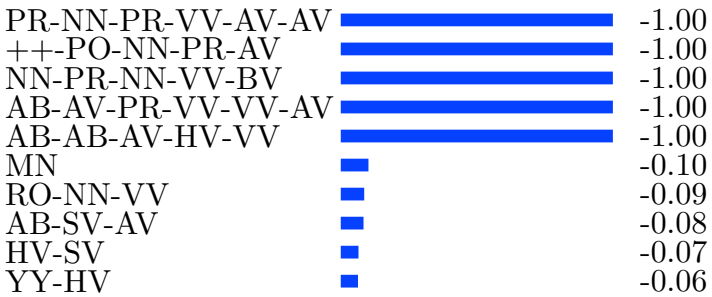


Figure 4.42: cluster A ⇔ cluster D, leaf-head features

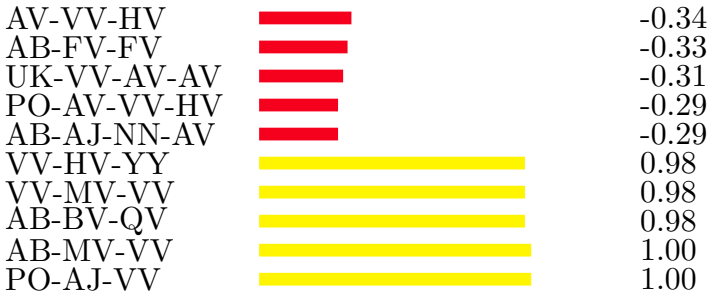


Figure 4.43: cluster B ⇔ cluster C, leaf-head features

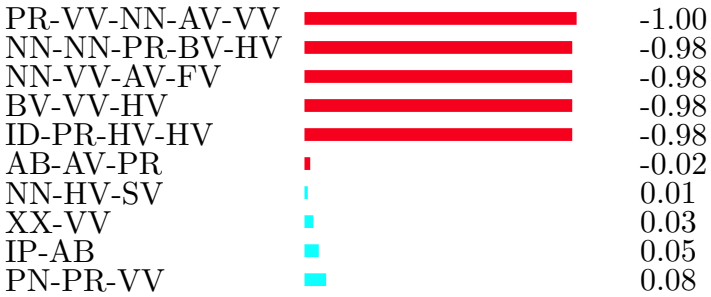


Figure 4.44: cluster B ⇔ cluster D, leaf-head features

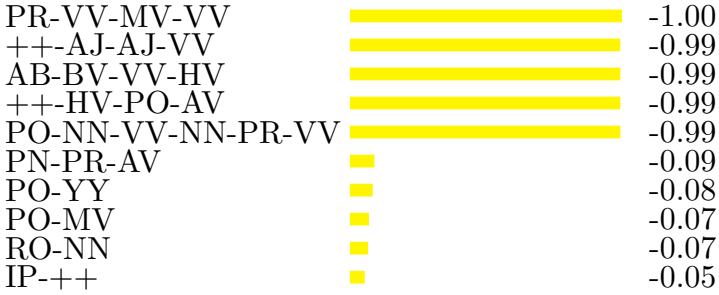


Figure 4.45: cluster C ⇔ cluster D, leaf-head features

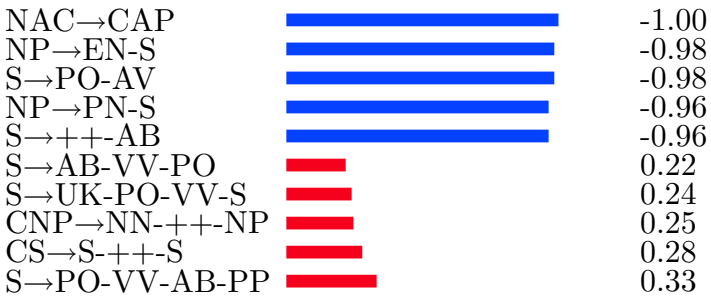


Figure 4.46: cluster A ⇔ cluster B, phrase-structure rule features

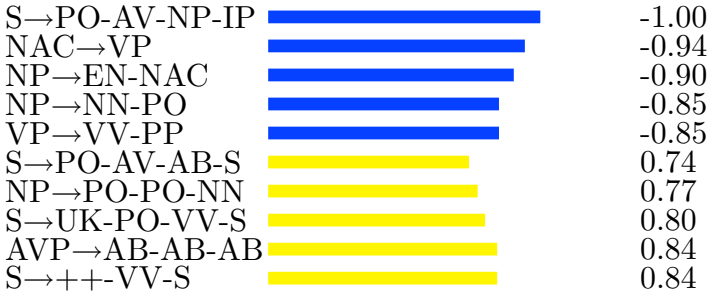


Figure 4.47: cluster A ⇔ cluster C, phrase-structure rule features

Overall both normalizations leave something to be desired; without overuse normalization, only very common features appear. These features convey only basic information, making it hard to identify characteristics of a cluster. On the other hand, the overuse normalization is susceptible to noise, especially for more error-prone feature sets. Even though more detail may be available with this normalization step, the features must be inspected for general trends because individual features are not necessarily reliable.

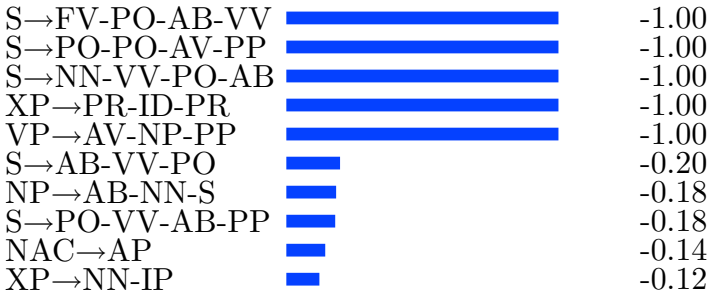


Figure 4.48: cluster A ⇔ cluster D, phrase-structure rule features

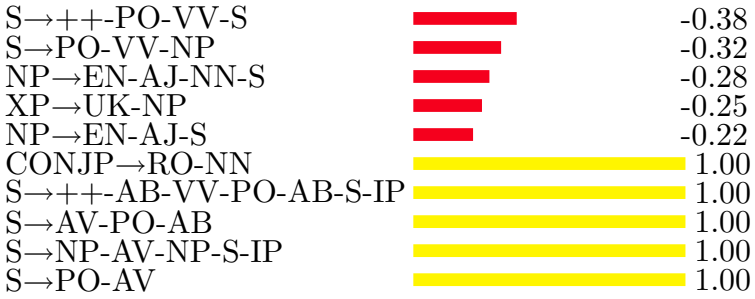


Figure 4.49: cluster B ⇔ cluster C, phrase-structure rule features

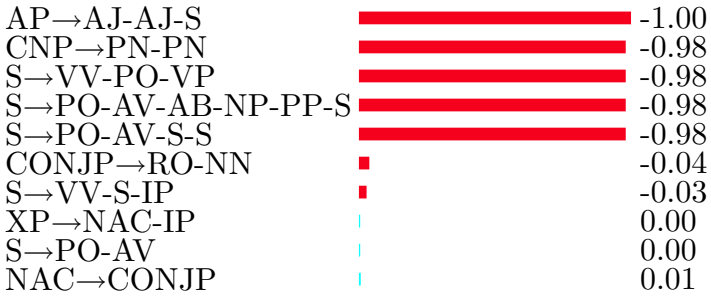


Figure 4.50: cluster B ⇔ cluster D, phrase-structure rule features

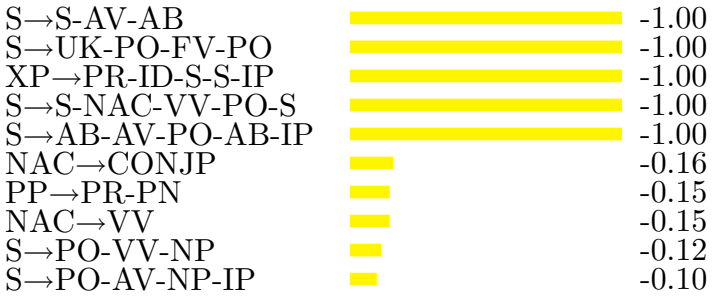


Figure 4.51: cluster C ⇔ cluster D, phrase-structure rule features

4.8 Conclusion

This chapter provided results and analysis of the results with no comparison to other work. The next chapter will compare the results to dialectology, phonological dialectometry, and previous work in syntactic dialectometry. Even before this comparison, however, the distance measure seems to be successful at producing dialect distance.

Quite a few patterns are visible: the significance tests show that the distance measure is finding significant distances for most parameter settings; the analysis of correlation shows that some correlate with geographic and travel distance. The dendrogram maps, composite cluster maps, and MDS maps all show a picture of with fairly well-defined areas. Finally, the feature rankings show some interesting patterns but nothing definitive. More qualitative analysis is needed in the future.

5

Discussion

This chapter compares the dissertation's results to three areas. It compares the results to dialectology, starting with the traditional dialect regions of Sweden and moving to individual dialect phenomena. Then it compares the results to phonological dialectometry, which uses many of the same analytical techniques on phonology data. Finally, it compares this work to previous work in the field of syntactic dialectometry and summarizes its improvements. The chapter ends with a summary of the work and its contributions to dialectology at large and Swedish dialectology in particular.

5.1 Comparison to Syntactic Dialectology

The comparison to syntactic dialectology consists of three sections. The first section looks at the general expectations of dialectology with respect to correlation with geographic distance. The second section compares the traditional dialect regions of Sweden to the ones found by the statistical dialect measure. The third section finishes by comparing specific phenomena of dialect regions to the corresponding features from interview sites.

General Expectations

The default expectation of dialect distance is that it should correlate with geographic distance, see Chambers and Trudgill (1998) and Gooskens (2004). The principal places that geographic distance fails to correlate with dialect distance are where dialect boundaries that exist between adjacent sites; here, a small geographic distance is paired with large dialect distance. For non-adjacent sites, in contrast, a large geographical distance may be paired with a small dialect distance. This can occur, for example, with relic dialects, where the innovative dialect expands from the center, leaving similar dialects isolated on the edges. However, neither of these cases holds for the Scandinavian languages; Hallberg (2005) points out that Swedish dialect areas form a continuous gradient without any strong boundaries. This means a particularly strong correlation between geography and dialect. Therefore, the first step is to compare the correlation of geographic distance with dialect distance as measured here.

Unfortunately, the correlations between geographic distance and dialect distance are uniformly low, even when they do attain significance. The highest correlation is 0.36. Correlating dialect distance with travel distance rather than geographic distance gives 0.37, which is an improvement, albeit a small one. However, as Gooskens point out, time and distance required to travel between two points at the beginning of the 21st century is considerably less than it was one hundred years ago or more. Measuring travel time between sites at some point in the past as she does might provide an even better correlation with dialect distance.

Nonetheless, the overall pattern agrees with Hallberg's analysis; there is a north-to-south gradient that is fairly smooth; the composite cluster maps (figure 4.18 in chapter 4, for example) show this pattern best, but the consensus tree and MDS maps do as well. The exceptions to this gradient are the areas surrounding Stockholm and Malmö, as well as the whole of the southern provinces Skåne and Blekinge. It may be that modern urbanization has created a city/country divide, with Stockholm and Malmö innovating and the rural areas becoming relic dialects. These two exceptions will be discussed more in the next section.

Dialect Regions

According to dialectology, Sweden does not have strong dialect boundaries, but it still has some traditional dialect areas. However, these are loosely defined and do not have sharp borders; the Eastern area is centered around Stockholm, the Western around Göteborg, the Southern around Malmö, and the Northern area covers the north of Sweden. In addition, the island of Gotland forms a separate area. The MDS maps and consensus tree maps reproduce these areas with varying degrees of fidelity.

For example, in the consensus tree figure 4.10, the cyan cluster corresponds to the Northern and Western dialect areas, the orange cluster corresponds to the Eastern area, and the red/yellow cluster corresponds to the Southern. There is a question that arises from this grouping, though; why should the northern and western areas appear in the consensus tree as one group? It looks as if the consensus tree map makes it more important that they differ from the East and South than that the differ from each other. The MDS maps reinforce this point; they show that the western sites and northern sites do in fact differ quite a bit. However, because the eastern and southern sites are so close, a clustering technique, like consensus trees, with exclusive group membership will put distant sites in the same group.

The boundary between the Skåne and Blekinge is quite abrupt, presumably mirroring the former Danish border that existed until the end of the Middle Ages. This contradicts Hallberg, who explicitly mentions that dialectology research finds no border there, and that the strongest north/south division more closely approximates Leinonen's (2008) diagonal boundary in map 5.18 below.

There are three possible explanations for this: first, there could be statistical, accumulative evidence which Swedish dialectologists have missed; second, the distribution of Swediasyn interview sites may be too sparse to reflect the real border; in particular, there are very few sites in Småland; third, the dialect landscape may have changed since the prevailing dialectology opinion was established. The last explanation is attractive, since the Swedia corpus is around 50 years newer than newest dialectology studies. However, this is an old boundary: it mirrors the Sweden-Denmark political border that existed over 400 years ago. It would be odd for it to disappear for over 350 years

and re-appear just before 2000. Instead, I believe the first explanation is more likely: Leinonen's results, in addition to reproducing the boundary described by Hallberg, also place a boundary at the same location as these syntactic results. This boundary is visible in factors 2 (figure 5.17) and 5 (figure 5.19) of her factor analysis based on the phonology data of Swedia, discussed below in section 5.2. Both Leinonen's method and mine are capable of detecting distributional patterns that are difficult to see from manual analysis. For example, in the previous chapter, I showed that the trigrams AB-AV-AB, despite appearing in all interview sites, was more common in central Swedish cluster A.

Dialect Features

The literature for Swedish syntactic dialectology is not extensive, largely because there is not much syntactic dialectology for any language. As a result, I will compare my results to two papers, Delsing (2003) and Rosenkvist (2007). The first paper is a survey of syntactic dialectology from the late 19th and early 20th centuries. In the same volume, other papers analyze specific phenomena in more detail; the survey is mostly concerned with the dialect differences and distributions rather than the syntactic analysis. The second is an analysis of the South Swedish Apparent Cleft.

Delsing (2003) surveys a number of dialectology studies. These studies date from the height of the field in Sweden, from circa 1880–1930, which Delsing at times augments with modern data. It is worth noting that the Swedia data in the comparison was collected around 2000, so there were likely changes in the dialects in the intervening 70–120 years. This is particularly true in the northern dialect areas, where improved travel and communication have leveled the dialects considerably (Hallberg 2005).

However, comparison to the phenomena in the survey may still yield interesting results, so for each phenomenon I will start with a summary of the phenomenon for Swedish dialects: its geographic distribution and its linguistic realizations. Then I will match the geographic distribution with Swediasyn interview sites and represent the phenomenon in terms of the feature sets developed in this dissertation. For this initial analysis, trigram features are used because they are simple. This matches Delsing's survey descriptions, which are for the most part surface-oriented;

Hä finns vattne däri hinken.
Here found water-the in bucket-the

'There is water in the bucket.'

Figure 5.1: Suffix marking for partitive

other papers in the same volume with his survey analyze the phenomena in more detail.

With the target sites and features defined, it is straightforward to count the number of occurrences of each feature in each site and compare the two. If the predicted dialect phenomenon is reflected in the data, then the sites associated with the phenomenon will have more occurrences of the target features than the non-associated sites. This difference is precisely what the distance measures use.

This method is inadequate for two reasons: first, the translation of linguistic analysis to feature representation will not be perfect and may miss some valid instances of the linguistic phenomenon. Second, more importantly, the differences are not yet checked for statistical significance. As such, the comparison can only be suggestive; checking for statistical significance will have to wait for future work.

The maps reproduced here are taken from Delsing's survey.

"Partitive" Article

Northern Sweden uses the suffixed article much more than the rest of Sweden. The reason, Delsing says, is that some uses of the suffixed article are not definite in the north; they have a partitive function, similar to the partitive article in French, which is not present in the rest of the country. See figure 5.1 for an example.

Unfortunately, the part-of-speech tag set used for this dissertation is quite coarse; it does not record whether nouns are marked with the definite suffix. Therefore, there is no way for the distance measure to tell the difference between suffixed dialect usage and bare standard usage.

Proper-Noun Articles

In Northern Scandinavia, first names are preceded by an indefinite article, and sometimes last names as well. The indefinite article also precedes kinship terms that are used as proper names, for example "Mother" or "Grandfather". An example is given in figure 5.3. Standard Swedish does not include this feature. In Sweden, this feature is found along the border with Norway as well as Northern Sweden. In the Swediasyn data, this includes the interview sites Köla, Indal, and Anundsjö—the dark area in figure 5.2, there labeled "Prepropriell artikel".

Unlike the partitive article suffix, this feature is easy to detect with a coarse part-of-speech tag set. Specifically, it can be represented as the bigram EN-PN (indefinite article-proper noun), which can be used as a search term in the trigram feature set. The same EN-PN sequence is expected for leaf-head paths, since the indefinite article depends on proper noun. The phrase-structure-rule features should look something like NP→EN-PN.

Occurrences of the EN-PN bigram in the trigram feature set for Leksand, Indal and Köla agree with the linguistic analysis: a rate of 0.00007 versus 0.00006. Unfortunately, this result cannot be trusted because the rate of occurrence for both regions is so rare, as well as so close between the two regions. The only conclusion that can be drawn is that the hypothesis is not yet disproved.

Possessives and the article

In Swedish, and in the other Scandinavian countries, there is a good deal of variation in the handling of possessives with articles. In Swedish, normally only one is allowed in a noun phrase: either a possessive or a determiner, but not both. However, in Danish and the Danish-influenced areas of Sweden, both are allowed in certain cases: for example, when the possessive and determiner are separated from the noun by an adjective. Delsing gives an example from Danish, shown here in figure 5.5. This pattern also exists in the southwest corner of Sweden, very near to Denmark. In figure 5.4, this area is shaded left-to-right diagonally; it includes the interview site Bara. In addition, this pattern alternates with the standard Swedish pattern on the island of Gotland (cross-hatched on the map), which includes the interview sites Fole, Fårö and Sproge.

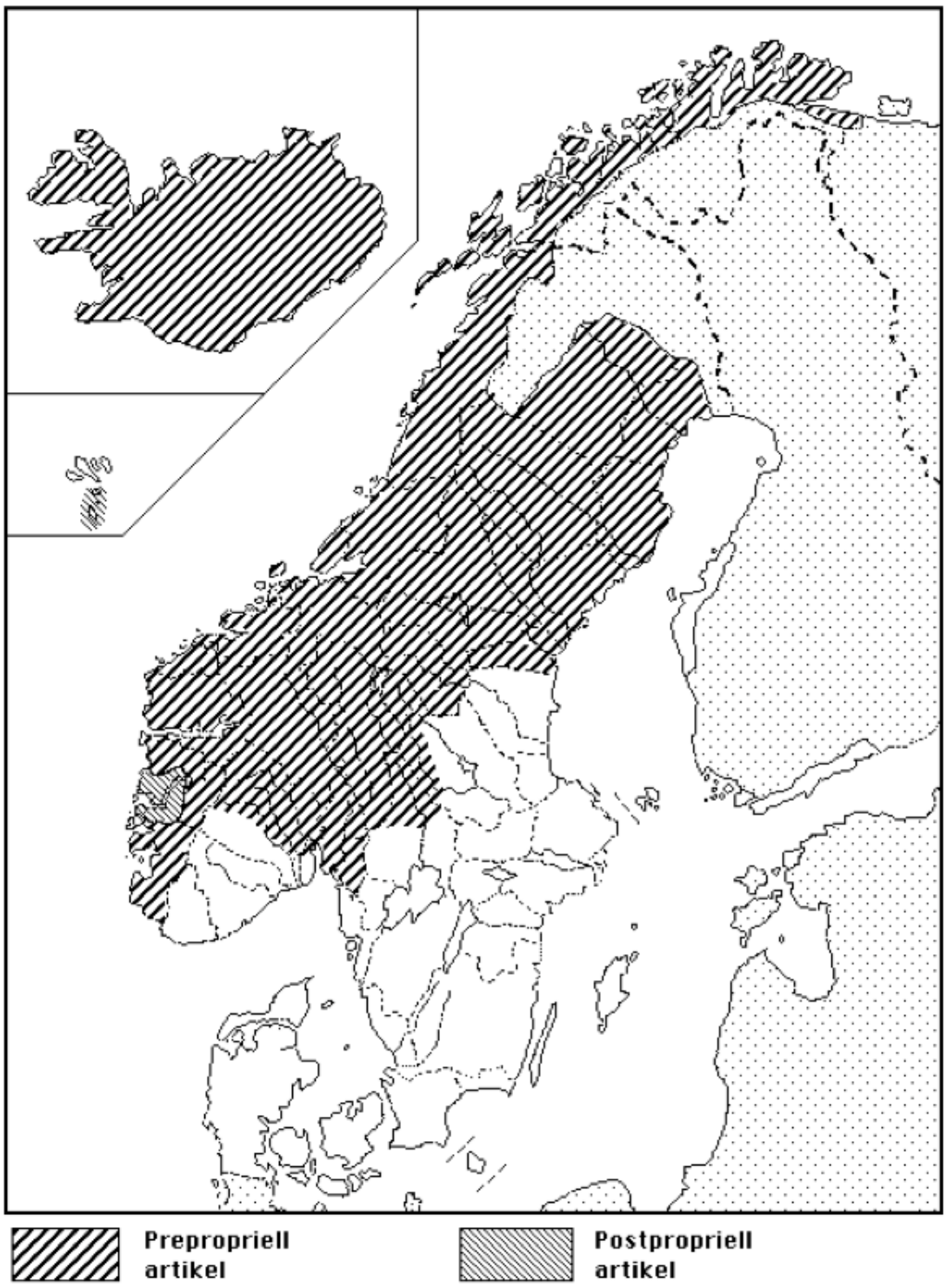


Figure 5.2: Proper-Noun Articles

En Bjurström ha affärn.
A Bjurström has the-store.

'Bjurström has a store.'

Figure 5.3: Indefinite Article for Proper Nouns: First Names

This pattern can be detected by analyzing the per-site recall for the 4-grams PO-PO-AJ-NN, PR-PO-AJ-NN and NN-PO-AJ-NN. The first is the sequence pronoun-pronoun-adjective-noun, for example *mitt det gamla huset* "My the old house-the". The second starts with a proper name, such as *Pers* "Per's", and the third starts with a noun, such as *naboens* "neighbor's". These three 4-tag sequence can be encoded as trigrams by breaking them into two pieces. This allows them to be searched for in that the distance measure would have encountered them.

In addition to this pattern, there is a second in the north of Sweden. Here, it is simply that possessive personal pronouns are allowed both before and after the noun. This pattern includes the interview sites Indal and Anundsjö and is covered in the next section.

Searching Bara, in the southwest of Sweden, for the previously mentioned trigram patterns does not find them: the rate of occurrence is 0.00289 inside Bara but 0.000341 outside. It should be higher in Bara. However, Delsing, writing in 2003, mentions that residents of Skåne that he has asked do not recognize this form either, so it is possible that it has fallen out of use in the 70 years or so since it was last reported.

Executing a similar search for the alternation of standard Swedish with the possessive pronoun pattern in the Gotland sites (Fårö, Fole and Sproge), the standard Swedish trigrams PO-AJ-NN, PR-AJ-NN and NN-AJ-NN show similar results: 0.00441 in Gotland, 0.00495 outside Gotland. This is opposite the predicted direction.

The final region in figure 5.4, in northern Sweden, which includes Indal and Anundsjö, is actually more complicated than can be captured by the part-of-speech tags used here; this region allows possessive proper nouns to occur with suffix-determiner nouns. But this can occur in either order: for example, both "Pers huset" and "huset Pers" is allowed. Although both "Pers hus" and "Pers huset" produce identical tags (PN-NN), trigrams do encode order, so the unusual order in "huset

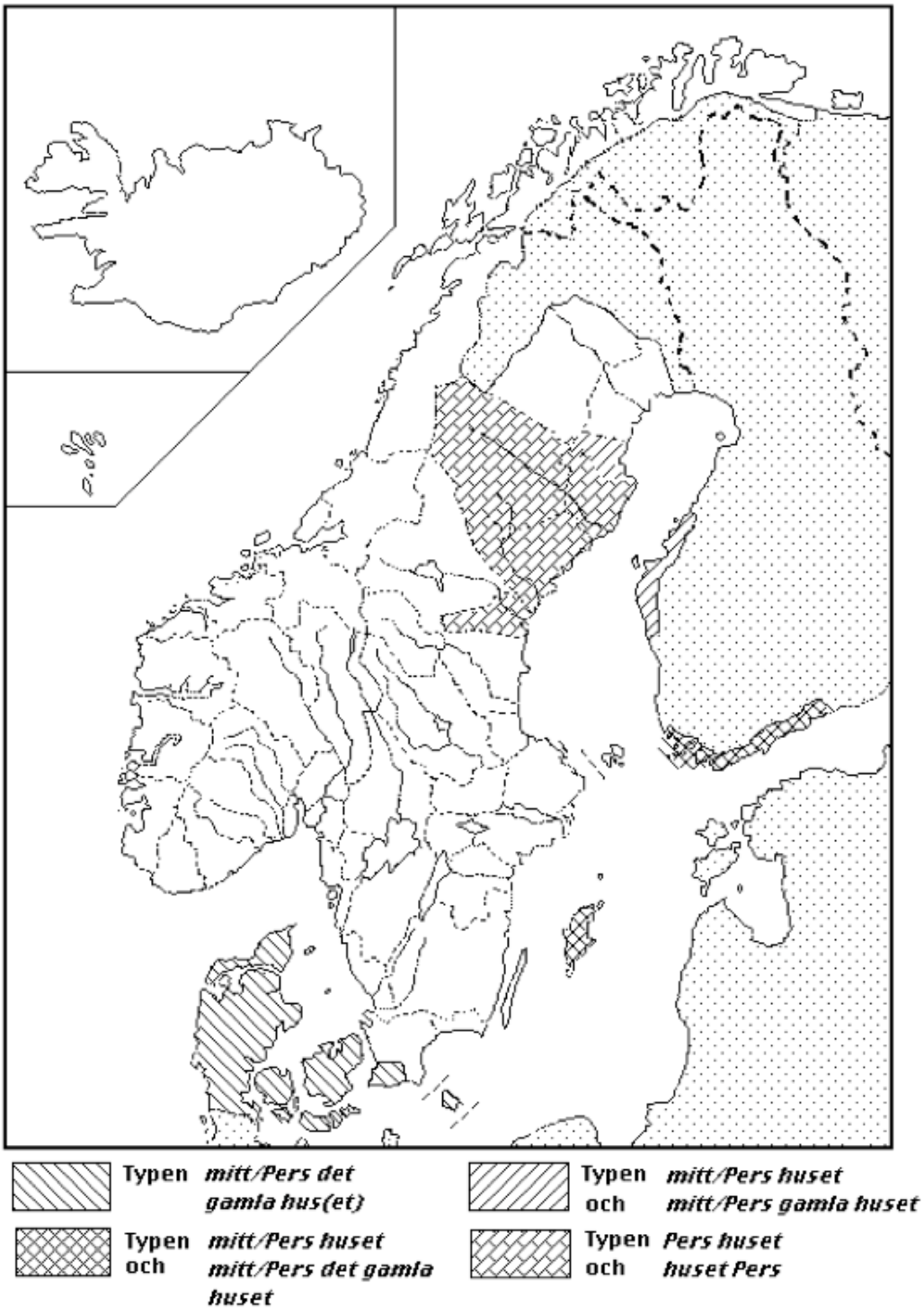


Figure 5.4: Proper-Noun Articles

naboens den stribede kat
Neighbors' the striped cat

'The neighbors' striped cat'

Figure 5.5: Simultaneous possessive and determiner in noun phrase in Danish, and at one time Southwest Sweden

Huset hans Per
The-house his Per

'Per's house'

Figure 5.6: Possessive formed of Possessive Pronoun and Proper Noun

Pers" can be searched for. Since both orders should be present in this northern area, it should overuse bigrams like NN-PN (noun-proper noun) relative to the rest of Sweden.

Searching for the bigrams NN-PN (noun-proper noun) and NN-PO (noun-pronoun) shows a usage rate of 0.02532 for Indal and Anundsjö and a rate of 0.02438 for the rest of Sweden. This is the expected direction, but the rate of usage is very similar between the two regions. The comparison is really too close to make a prediction because the difference is not likely to be significant.

Proper Noun Possessives

In addition to the post-nominal possessive pattern of the previous section, there is a variant that is common in Norway. Here, the sequence is noun-possessive pronoun-proper noun. An example of this pattern is given in figure 5.6.

This pattern overlaps slightly into Sweden, covering the interview site Köla. The distribution is given in figure 5.7. Note that the northern area with small stripes is the same as in figure 5.4, and the northern area with thin stripes has no matching sites. The area of interest is the one with larger, thick stripes that covers the majority of Norway.

This phenomenon maps to a trigram NN-PO-PN: noun-pronoun-proper noun. The occurrence rate of this trigram in Köla to the rest of Sweden is 0 vs 0.00001. This is the wrong direction, and the value is so low that it is probably noise. There are two possible causes for this essentially zero

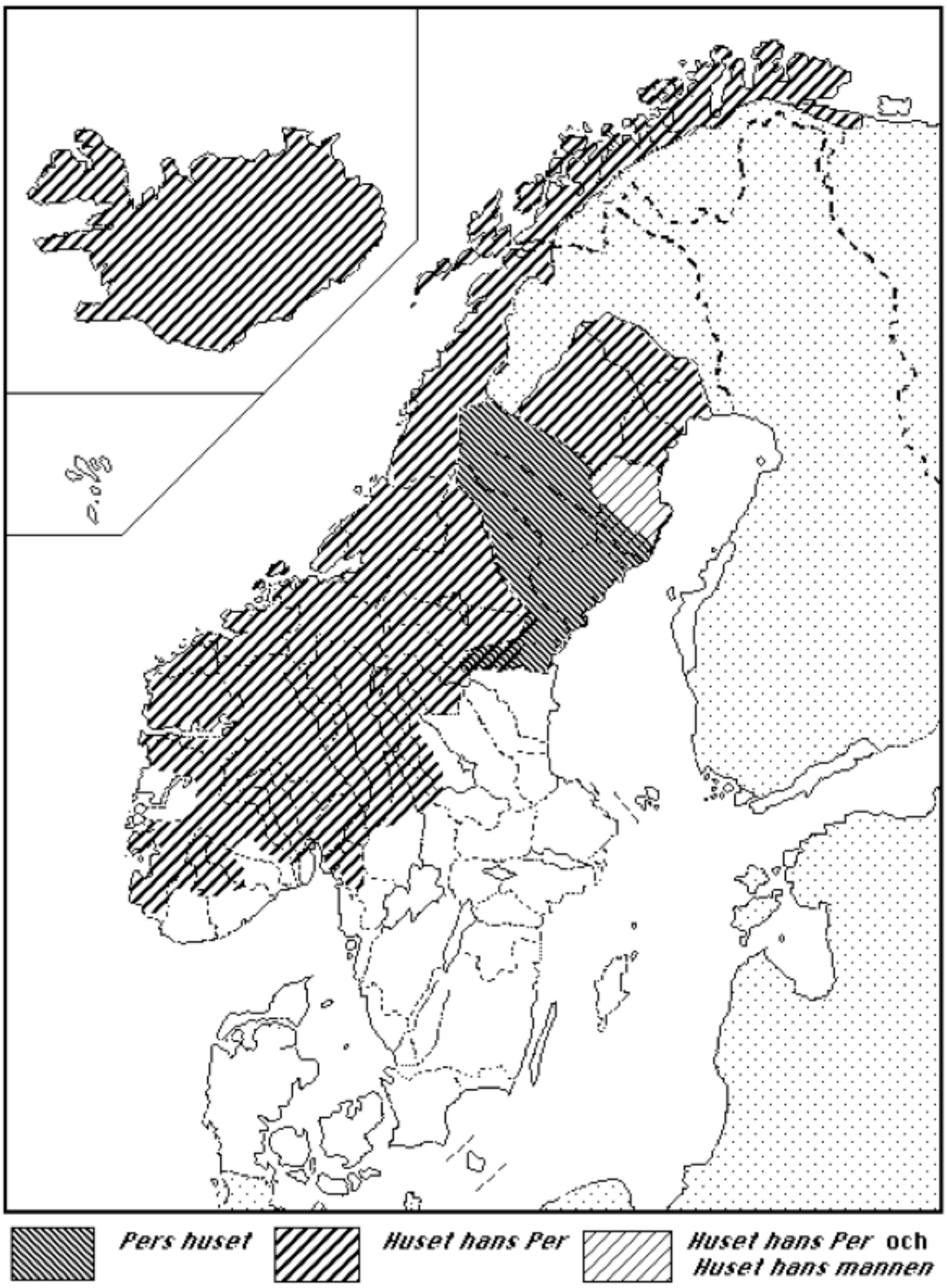


Figure 5.7: Proper-Noun Possessives

en stor en bil
a large a car

'A large car'

Figure 5.8: Double Indefinite

result: either neither region has this feature or there is not sufficient data to tell.

Noun possessives

Delsing mentions briefly that central Sweden, including Älvdalen and Västerdalarna, uses the dative form of nouns for the s-genitive. However, the part-of-speech tag set used here does not distinguish between dative and other cases on nouns, so it is not possible to represent this phenomenon in a way that the distance measures could have used.

Double indefinite

In northern Sweden and northern Norway, indefinite articles are used both before and after adjectives when modifying nouns. In map 5.9, this is the area covered by dark diagonals, labeled "Postadjektivisk artikel". Delsing also calls this the "double indefinite"; for an example, see figure 5.8. One indefinite article is used after each adjective, even for multiple adjectives, so *en stor en bil* (a large car) but also *en stor en fin en bil* (a large fine car).

In central Sweden, a similar pattern occurs, but the article is not perceived as independent. Instead it is perceived as a suffix of the adjective. In other words, the above example is perceived as *en stor-en bil* instead. According to Delsing, there is a difference in intonation compared to the North Swedish construction, which does not stress the intermediate articles nor co-ordinate them morphologically as would be expected with a suffix. Unfortunately, this pattern appears identical to the ordinary Swedish case given the course part-of-speech tag set in use. In contrast, the first pattern is quite easy to represent with trigrams: the 4-gram EN-AJ-EN-NN and the 6-gram EN-AJ-EN-AJ-EN-NN—alternating series of indefinite articles and adjectives ended by a noun. These

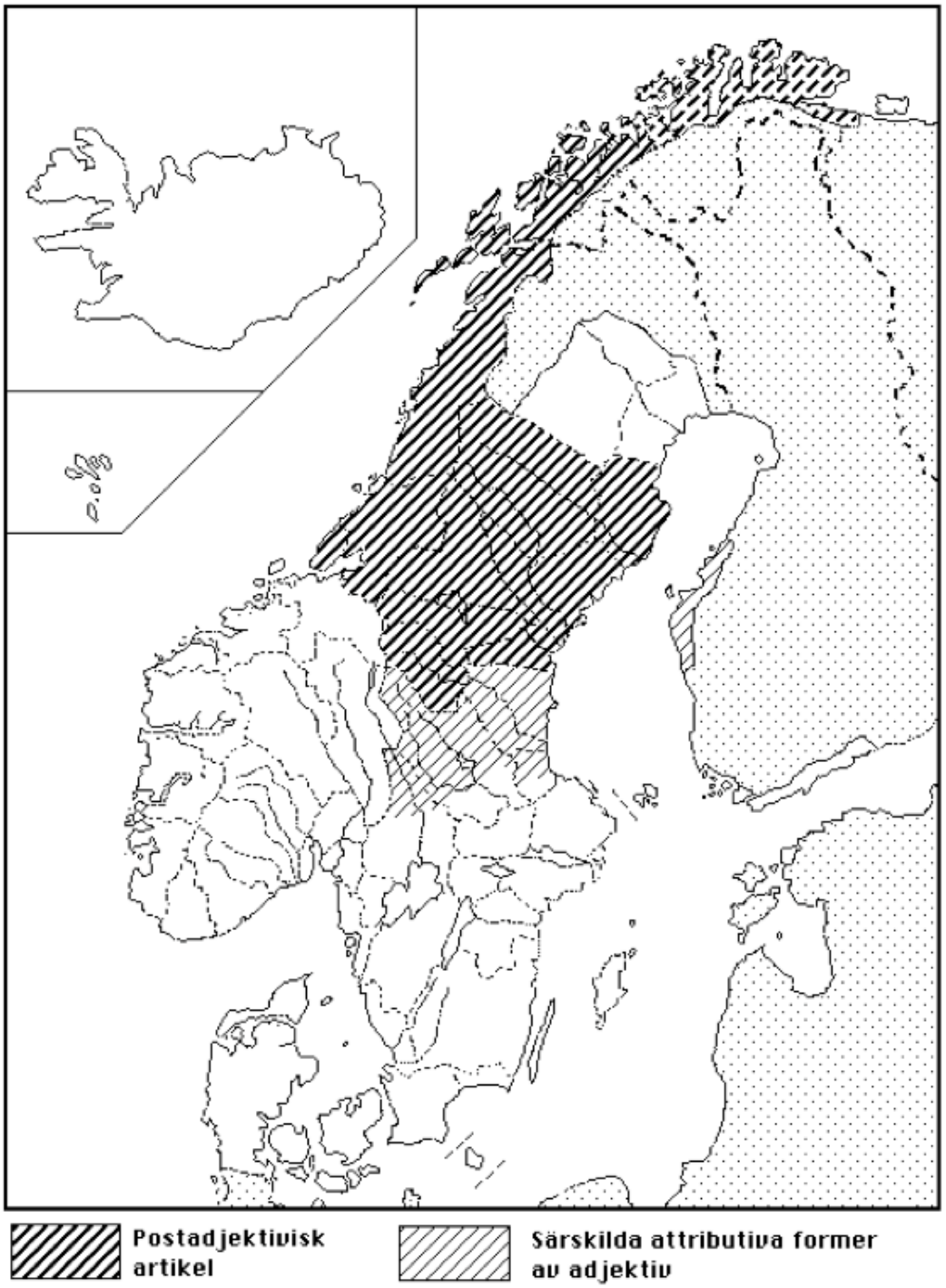


Figure 5.9: Double indefinite (post-adjectival articles)

det store huset
The large the-house

'The large house'

Figure 5.10: Double definite (Sweden and Norway)

det store hus
The large house

'The large house'

Figure 5.11: Single Indefinite (Denmark)

larger n-grams can be broken into the trigrams EN-AJ-EN and AJ-EN-NN in order to search for them in the Swedia-based data.

The northern pattern includes the interview sites Anundsjö and Indal. When measured, these trigrams occur at a rate of 0.00054 there versus the rest of Sweden, which has a rate of 0.00012. From this we can conclude that this is a rare phenomenon, but one that happens in the north about 4 times more often than in the rest of Sweden.

Double Definite

Double-definite with adjectives is standard in Sweden and Norway, where there is a definite article as well as a definite suffix on the noun (see figure 5.10). This is not the case in Denmark (figure 5.11), where the definite suffix disappears in case of a definite article, nor in Iceland, where the definite is suffix-only and there is no article (figure 5.12).

However, in North Sweden, there is a fourth option, where the adjective combines with the

gamla húsid
old house-the

'The old house'

Figure 5.12: Single definite suffix (Iceland)

storhuset
large-house-the

'The large house'

Figure 5.13: Single definite suffix with combined adjective (Northern Sweden)

noun into a single word (figure 5.13). Delsing gives examples like *storhuset* (the big house) and *stor-svart-gamm-katta* (the big, black, old cat), in which a series of adjectives appear prefixed to a noun without their usual morphological inflection. In Norrland, Delsing finds that this construction is used almost to the exclusion of the normal Swedish one. Further south, the two co-exist.

Therefore, since the annotation scheme does not differentiate between a combined noun like *storhuset* and a normal noun like *huset*, the better way to detect the region difference is to count the rate of normal trigrams like PO-AJ-NN (pronoun-adjective-noun); this is the feature type that occurs rarely or not at all in the north. If the region division in map 5.14 is detected, then northern Sweden will have a lower rate of occurrence of these standard trigrams.

As before, the two northern sites are Indal and Anundsjö. The rate of PO-AJ-NN in this region is 0.00152, compared to 0.00216 for the rest of Sweden. This difference is in the right direction, and it is larger than most of the other comparisons here. However, like the other comparisons, it has not been checked for significance so it is currently only suggestive.

Rosenkvist's Analysis of the South Swedish Apparent Cleft

Rosenkvist (2007) analyzes a phenomenon he calls the South Swedish Apparent Cleft. It involves an embedded clause, similar to a cleft, but with no clefted constituent. Instead, the subordinating conjunction *som* is directly preceded either by the verb or an adverb expressing speaker attitude. The subject of of the *som*-clause must be a pronoun, though Rosenkvist notes that this may be a pragmatic, not a syntactic, restriction. The two main variants are given in figures 5.15 and 5.16, but the apparent cleft is also found in yes/no questions and embedded clauses.

Unfortunately, Rosenkvist does not give a comprehensive syntactic analysis of the apparent cleft. This means that a translation to our feature set based on his description will necessarily be

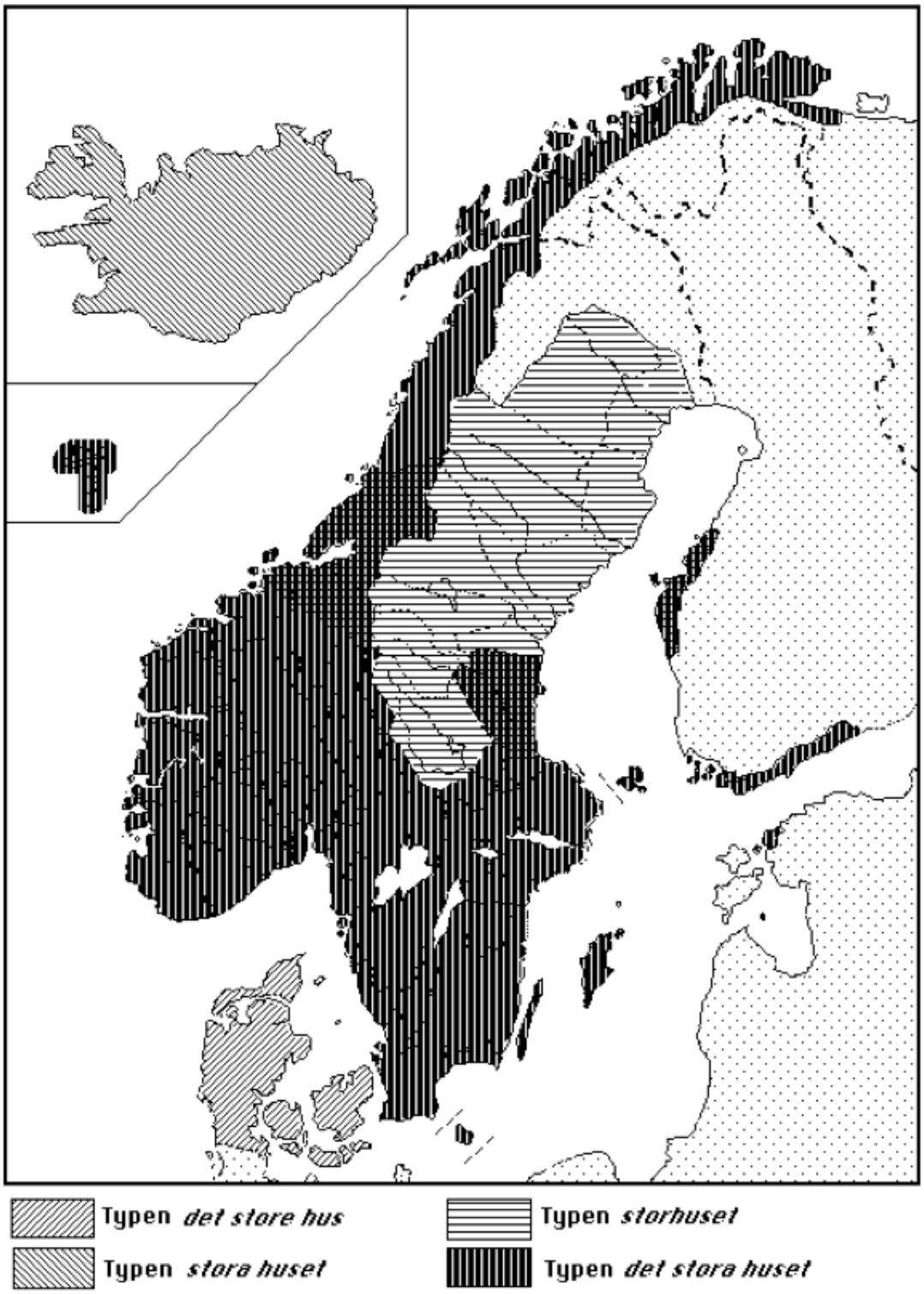


Figure 5.14: Double definite (and combined adjectives)

Det är som han har missuppfattat.
it is *som* he has misunderstood

'He has misunderstood.'

Figure 5.15: Apparent Cleft

Det är bara som han finner på.
it is only *som* he finds-on

'He just makes it up.'

Figure 5.16: Apparent Cleft with adverb expressing speaker attitude

surface-oriented in the same way this his analysis and results are surface-oriented.

Accordingly, translating the sequences like *Det är som han ...* gives the 4-gram PO-AV-UK-PO, and *Det är bara som han ...* gives the 5-gram PO-AV-AB-UK-PO (pronoun-be verb-adverb-subordinating conjunction-pronoun). Although these part-of-speech sequences can obviously appear in other contexts, they should appear more in the region that has apparent clefts than in the region that does not. Converting these sequences to trigrams is straightforward, producing 5 unique trigrams of interest, which the distances measures should also have used to obtain their distances.

Rosenkvist captures the geographical distribution of the apparent cleft in two ways. He first consults two collections of Swedish novels, using the authors' birthplaces as proxies for their dialect. Second, he uses the results of a questionnaire that he issued to university students at several Swedish universities: Stockholm, Gothenburg, Lund and Umeå.

Using author birthplace as a proxy for dialect, the apparent cleft can be seen throughout southern and middle Sweden—this includes all the interview sites except Årsunda, Indal and Anundsjö. However, based on the survey results, the apparent cleft is only accepted by speakers from Halland, Småland and Skåne. This includes the interview sites Frillesås, Våxtorp, Ankarsrum, Torsås, Bara, Löderup, Norra Rorum and Össjö.

Therefore, the test for this comparison is the occurrence rates for the 5 trigrams based on the two common forms Rosenkvist gives as examples, with two variations: one region division based on author birthplaces and one region division based on the student survey. The southern region in

both cases should have more occurrences of the target trigrams.

For the larger cleft region division based on author birthplaces, the comparison goes in the expected direction: a rate of 0.02430 in the south and 0.02427 in the north. But these rates are so close to identical that they should not be regarded as different. For the smaller division based on the student survey, the comparison goes in the opposite direction: 0.02264 in the south and 0.02491 in the north. Again, this is not much of a difference.

With such a small difference, it is not possible to draw any conclusions or even suggest whether the distance measures consistently notice this difference. One problem is that it is hard to capture a phenomenon like this with trigrams, where the surface form is only subtly different from that produced by other syntactic structures. A more complete syntactic analysis of the phenomenon is needed so that more advanced feature sets from dialectometry can be used to compare to the results from dialectology.

Conclusion

The dialect constructions surveyed here do not support the agreement of the new dialectometry results with existing dialectology results nearly as well as the previous sections which compared the results at a less detailed level. The larger problem is that no good method yet exists for doing so; the differences were in some cases large enough to be suggestive, but without significance testing, it is not possible to know that they are reliable. It is possible that the small differences are significant, and already being used by the distance measures to distinguish regions; after all, the aggregation of many small differences is the inherent in the working of the statistical approach in this dissertation.

5.2 Comparison to Phonological Dialectometry

The comparison to phonological dialectometry is currently difficult in two ways. First, there are few statistical methods in phonological dialectometry. I proposed a simple Bayesian method (Sanders and Chin 2006) and Hinrichs and Zastrow (2007) proposed two more complex methods,

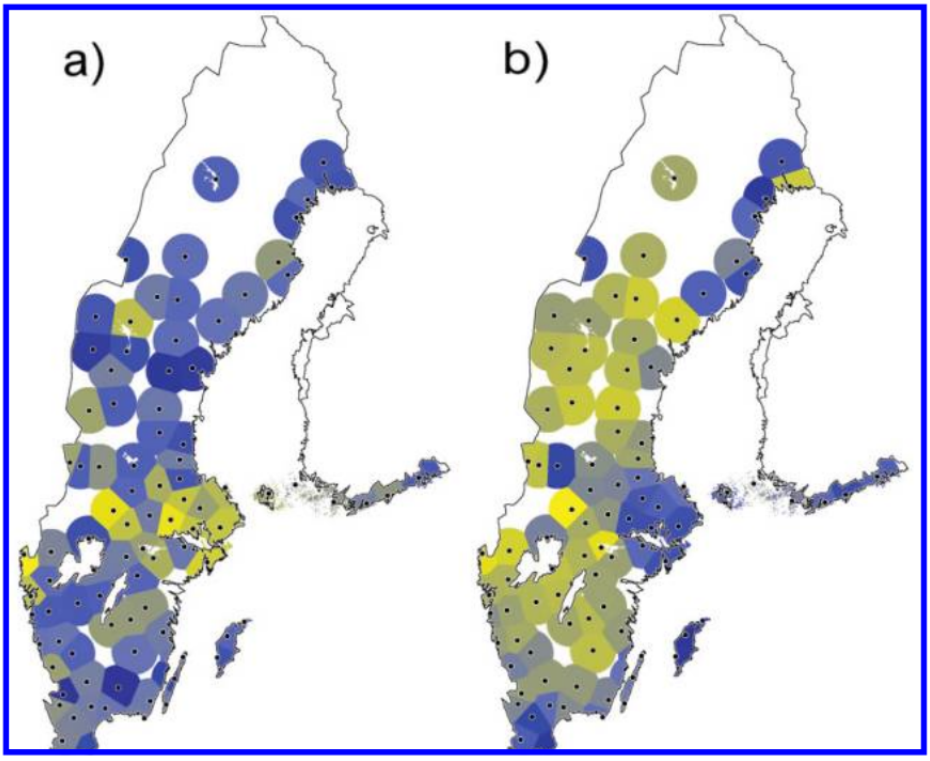


Figure 5.17: Factors 1 and 2 of Swedish vowels

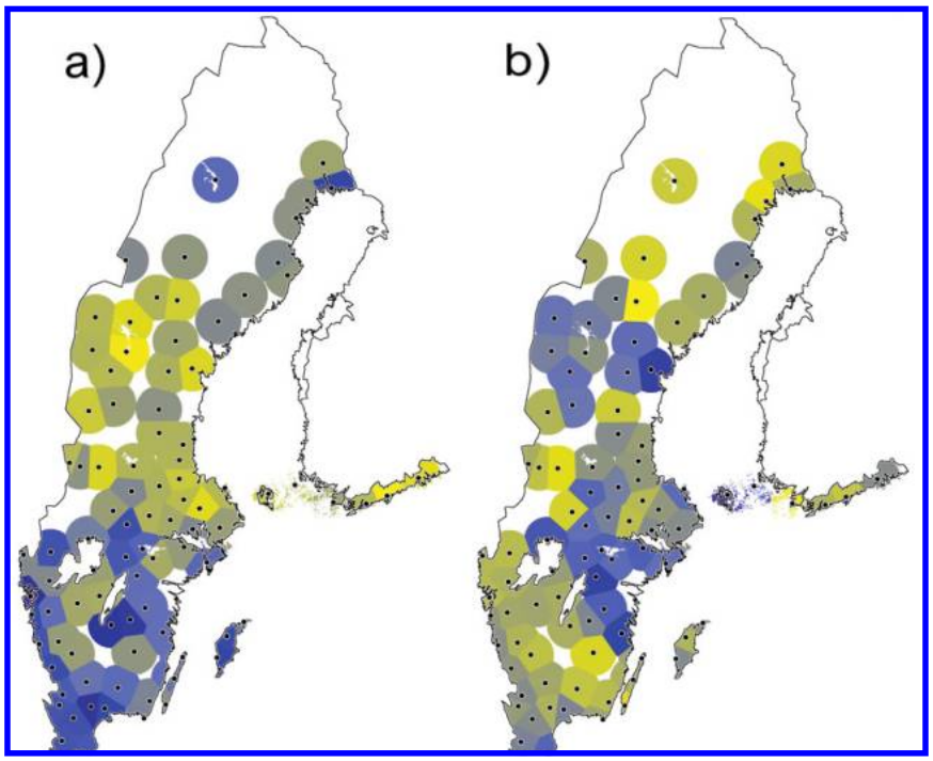


Figure 5.18: Factors 3 and 4 of Swedish vowels

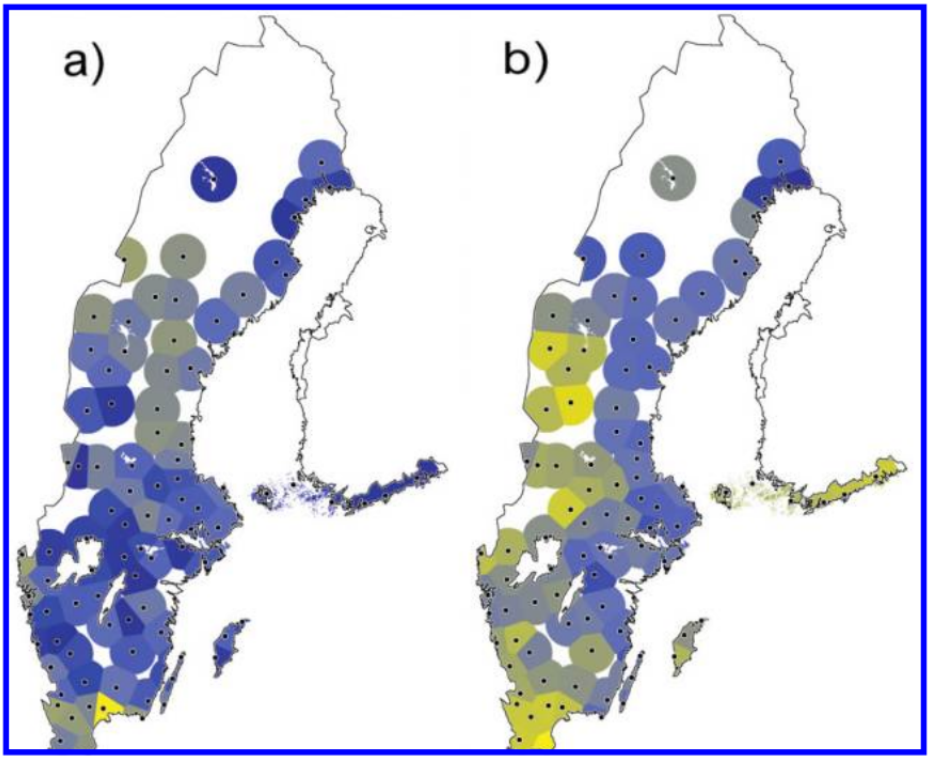


Figure 5.19: Factors 5 and 6 of Swedish vowels

one vector-based and the other from information theory. However, these methods are less effective on small corpora than Levenshtein distance and have not gained traction in the field. Second, even comparing results only, there has been little Swedish dialectometry to date. To my knowledge, the only paper at the time of this writing is Leinonen (2008); its method is more similar to Spruit's (2008) approach to syntax. It uses factor analysis to characterize the distribution of nine phonological variables across Sweden, but does not cluster the sites based on these variables. However, the overall regions can still be compared. I compare Leinonen's individual feature maps to my composite cluster and MDS maps.

In addition, Leinonen's dissertation, currently unpublished, will cover phonological dialectometry of Sweden comprehensively. In future work, a better comparison should be possible, since both dissertations are based on the same corpus.

Looking at Leinonen's first two maps, reproduced here as figure 5.17, we see patterns similar to the city/countryside difference from the syntactic results: in the first diagram, Stockholm and Uppsala differ from the rest of the country, and in the second Stockholm, Uppsala and Malmö areas all differ.

In Leinonen's third and fourth maps (figure 5.18), there is a north/south divide roughly half way between Stockholm and Malmö. This boundary generally reflects the north/south gradient from my results. However, the phonological boundary is stronger and more localized than numerous small syntactic ones, such as those seen in the composite cluster map 4.18. It is closer to the diagonal north/south boundary mentioned by Hallberg (2005).

The fifth map (figure 5.19) is more specific than the previous four; most of the sites are blue, but there are a few in the south that are much yellower than the rest. These are the same three sites that form the red cluster in figure 4.12 from the consensus tree results in chapter 4: Jämshog, Össjö and Torsås. The sixth map, however, shows a clear east/west divide that is not reflected in my data.

Although this region-to-region comparison is not precise, it provides hope that a quantitative comparison between the two result sets will support high agreement with statistical evidence. The level of agreement between the phonological results and syntactic results is quite high. Of the six variables Leinonen illustrates with the maps in figures 5.19 – 5.18, all but one reflect some aspect

of the combined syntactic results. The exact overlap between Leinonen's fifth variable and the red cluster from the consensus tree results is surprising for statistical methods.

5.3 Comparison to Syntactic Dialectometry

In the progression from dialectology of Swedish to phonological dialectometry of Swedish and finally to syntactic dialectometry, there is less and less existing literature. To my knowledge, this dissertation is the first treatment of syntactic dialectometry for Swedish. Even outside Swedish, very little syntactic dialectometry exists. Besides Spruit's (2008) dissertation, based on Goebl's limited-data techniques, statistical work is limited to Nerbonne and Wiersma's work on Finnish (Nerbonne and Wiersma 2006) and (Wiersma 2009), and my work on English (Sanders 2007) and (Sanders 2008).

This dissertation is the first to show that a statistical measure designed for syntax can find distances between dialect regions. It directly addresses the shortcomings of the previous work, which showed that a statistical measure could detect significant differences, but failed to produce dialect distances. It evaluates parameter variations, establishing which combinations of feature set, distance measure and corpus size produce valid and useful results, taking into account a number of practical considerations, such as amount of existing annotation.

This dissertation shows that fairly small sites, on the order of 6,000–10,000 words, can produce significant distances. This contrasts with previous work; the significant distances between English sites were for much larger sizes: the ICE data for London had over 200,000 words, and Scotland over 25,000. The conclusion should be that when the sites consist of properly collected dialect speech, the size required to detect distance drops considerably. The Swediasyn corpus captures dialect speech in a way that the ICE does not; the Swediasyn contains interviews in homes, while the majority of the ICE is interviews of students and professors at University College London.

In addition, the syntactic results of this dissertation agree closely with the phonological results of Leinonen (2008). Although agreement of syntax and phonology is not necessarily a prediction when looking for dialect regions, it is not surprising—circumstantial evidence that a new method is

valid because it agrees with an existing one. This contrasts strongly with the English work, which found no significant correlation of syntactic distance with phonological distance. It may be that using the same corpus for both Swedish studies was the key difference; the two English corpora's ages differed by almost 50 years.

This dissertation agrees more closely with dialectology than previous work. Although the English study reproduced the north/south divide well known in British dialectology, it did not produce any more detailed regions. In contrast, this study reproduced all of the Swedish dialect regions. With respect to individual phenomena, however, the feature comparison was inconclusive; a few results were positive, but most were very close to zero. There are two problems: the corpora once again differ in age—most of Swedish dialectology dates from around 1900 while the Swediasyn was collected in 2000—as well as a lack of significance testing. The small feature differences found may well be significant, since the nature of statistical methods is to accumulate many small differences, but it is not possible to tell without a test.

Significance testing for precise feature analysis is future work, but this is not necessarily a problem. For phonological dialectometry, which began with Kessler's paper on Irish (Kessler 1995), extraction of specific features did not begin until much later, two to three years after Heeringa's dissertation on the subject (Heeringa 2004), with such work as Prokić's (2007). In any case, Wiersma (2009) mentions a method for features of individual regions that could be adapted to comparisons between a pair of regions.

6

Conclusion

The previous chapter discussed the impact of this work with respect to previous work in various fields. In particular, it provided a picture of how it advanced syntactic dialectometry. This chapter briefly covers avenues of future work to which this work leads. This future work falls into two categories: syntactic dialectometry and Swedish dialectology.

6.1 Future Work

Some avenues of future work are obvious; Swediasyn is part of the larger Nodalida project to create a syntactic dialect corpus for all Scandinavian languages. And Swediasyn is itself not a complete transcription of Swedia; for example, it does not include any of Swedish-speaking Finland yet (Johannessen et al. 2009). Unfortunately, this work depends on others since I do not speak any Scandinavian language natively. Once these corpora are complete, they will provide a more complete picture of syntactic variation over the entire Scandinavian language area.

With regard to feature sets, it is interesting that trigrams perform better than the more complicated feature sets. From a linguists' point of view, this is disturbing: why should the flattest representation of syntax perform the best? This performance difference also discourages others from developing even more complicated and linguistically interesting feature sets. The reason for trigrams' performance is likely because of the amount of automatic annotation that is a prerequisite for the complex features developed here. Trigrams rely on an automatic part-of-speech tagger,

while leaf-ancestor paths rely on an automatic parser that uses automatic part-of-speech tags from that same tagger.

To enable more complex feature sets, manual annotation is needed. But this is labor intensive. Failing that, improved automatic annotation is needed, although this still usually implies some manual annotation in the form of a seed corpus for bootstrapping (Blitzer et al. 2006, Blitzer 2007). Bootstrapping should help automatic parsing of dialect interviews: not only does the subject matter of an interview differ from the typical newspaper training corpus, the syntactic features where the dialect differs from the standard language are precisely those that are hardest to parse. Giving a machine parser a sample of dialect speech as training would allow it to identify some of these features. For example, in the case of the possible double modals discussed at the end of chapter 4, the part of speech tagger never saw the tokens *"skulla kunna"* juxtaposed in the training. If both words were not part of a closed class, it is likely that the tagger would not produce the correct tag for this pair. The same problem applies to parser, but because syntactic training is even more sparse, the parser is less likely to to have seen similar structures in training. The parser is correspondingly less likely to produce a double modal structure without having seen it in training.

Processing of features is another area for future work: normalization is the first half of this problem. The current sentence-level normalizations function well for aggregate comparisons like cluster maps, but for individual feature comparison, the overuse normalization tends to rank highly features that may just be noise from the annotation error. On the other hand, without the overuse normalization, only very common features are high ranked. This makes it hard to notice the unique features of a dialect that do not occur much. A compromise that takes frequency into account to some extent is needed, so that rare features can be highly ranked without introducing noise from annotation errors.

The other half of the feature-processing problem is a test for significance when comparing two regions. This would make sure that comparisons to the dialectology literature are significant in the future. Wiersma (2009) provides a similar method for testing significance of individual features in a single region, so it should be easy to modify this to work for comparisons between two regions.

Besides general improvements to feature-processing, many improvements are possible to the

individual feature sets used here. In order to compare the Berkeley parser to MaltParser, dependency parses should be extracted from its constituency parses. When extracting features from dependency parses, the current leaf-head paths contain either node labels or arc labels, but not both. As long as data sparseness is not a problem, leaf-head paths with both node and arc labels would capture more information than either alone. In addition, the compound feature set used in this dissertation is very simple: a linear combination of feature sets. A non-linear combination with trained weights could provide better performance, as could backoff from sparse feature sets to simpler ones. For example, backoff from node/arc-labeled leaf-head paths to node-labeled paths, or backoff from POS 4-grams to trigrams.

Finally, another obvious extension of this work is a quantitative comparison of these results on Swedish to the results in Leinonen's upcoming dissertation on Swedish phonological dialectometry. Given the agreement between these results and her published work, it is likely that the correlation will be high. This comparison should be fairly easy since both results use the same dialect corpus as a basis.

6.2 Conclusion

This dissertation establishes that statistical methods are useful direction for syntactic dialectometry. Its results show that significant differences can be obtained with dialect corpora. This much had been accomplished by previous work. However, this work goes on to establish that even smaller interviews of dialect speakers are sufficient to produce significant distances, and investigates variations on both feature set and distance measure. It shows that a syntactic measure can reproduce the traditional regions of dialectometry, and that it can produce agreement with a phonological measure. Its comparison to individual dialect phenomena is inconclusive, but opens an avenue for future investigation, and more importantly, future development of methods to compare and rank individual features.

Future directions based on this work are twofold. First, with a statistical method established for syntax, dialectometry can begin to investigate the syntactic features of other languages. Second, in

Swedish, this work and future work similar to it can contribute to dialectology in general; syntax has been relatively neglected in Swedish dialectology. As Swediasyn and Nodalida are completed, the automatic analysis detailed in this dissertation can provide a quick analysis of new data, and point linguists toward interesting dialect features.

In conclusion, this dissertation has answered the questions of agreement with dialectometry and best parameter configuration for practical measurements, as well as agreement with phonological dialectometry. It has established statistical methods for syntactic dialectometry, pointing the way for future syntactic dialect studies, future expansion of statistical methods in dialectometry, and future syntactic analysis of Swedish.

Bibliography

Amenta, Nina, Frederick Clarke, and Katherine St. John. 2003. A linear-time majority tree algorithm. *Lecture Notes in Computer Science* 2812:216–227.

Aron, Arthur, Elaine N. Aron, and Elliot J. Coups. 2006. *Statistics for Psychology*. Prentice Hall. Fourth edition.

Barbiers, Sief, Hans Bennis, Gunther De Vogelaer, Magda Devos, Margreet van der Ham, Irene Haslinger, Marjo van Koppen, Jeroen Van Craenenbroeck, and Vicky Van den Heede. 2005. *Syntactic Atlas of the Dutch Dialects*. Amsterdam University Press.

Blitzer, John. 2007. *Domain Adaptation of Natural Language Processing Systems*. PhD thesis, University Pennsylvania.

Blitzer, John, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 120–128.

Brants, Thorsten. 2000. TnT : A statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied Natural Language Processing*, 224–231, Seattle, Washington. Association for Computational Linguistics.

Bruce, Gösta, Claes-Christian Elert, Olle Engstrand, and Pär Wretling. 1999. Phonetics and phonology of the Swedish dialects – a project presentation and a database demonstrator. In *Proceedings of ICPhS 1999*.

Bryant, David. 1997. *Building Trees, Hunting for Trees, and Comparing Trees: Theory and Methods in Phylogenetics*. PhD thesis, University of Canterbury.

Chambers, J. K., and P. Trudgill. 1998. *Dialectology*. Cambridge, United Kingdom: Cambridge University Press. Second edition.

Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English*. Harper & Row.

Clopper, Cynthia, and David B. Pisoni. 2004. Homebodies and army brats: Some effects of early linguistic experience and residential history on dialect categorization. *Language Variation and Change* 16:31–48.

Collins, Michael, and Nigel Duffy. 2001. Convolution kernels for natural language. In *Advances in Neural Information Processing Systems 14*, 625–632. MIT Press.

Delsing, Lars-Olof. 2003. Syntaktisk variation i nordiska nominalfrasen. In O. A. Vangsnes, A. Holmberg, and L.-O. Delsing (Eds.), *Dialektsyntaktiska studier av den nordiska nominalfrasen*, 11–64. Novus Press.

Geršić, S. 1971. Die Berechnung der phonetischen Variabilität: ein Beitrag zum objektiven Vergleich phonetischer Texte. *International Congress of Phonetic Sciences 7* 110–111.

Goebl, Hans. 2006. Recent advances in Salzburg dialectometry. *Literary and Linguistic Computing* 21(4):411–436.

Good, Phillip. 1995. *Permutation Tests*. New York: Springer.

Gooskens, Charlotte S. 2004. Norwegian dialect distances geographically explained. In B.-L. Gunnarson, L. Bergström, G. Eklund, S. Fridell, L. H. Hansen, A. Karstadt, B. Nordberg, E. Sundgren, and M. Thelander (Eds.), *Language Variation in Europe. Papers from the Second International Conference on Language Variation in Europe*, 195–206, Sweden, June. ICLAVE 2 Uppsala University.

Hallberg, Göran. 2005. Dialects and regional varieties in the 20th century I: Sweden and Finland. In O. Bandle, K. Braunmüller, E. H. Jahr, A. Karker, H.-P. Naumann, and U. Teleman (Eds.), *The Nordic Languages*, Vol. 2, chapter 185, 1691–1706. Walter de Gruyter.

Heeringa, Wilbert J. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. Doctoral dissertation, University of Groningen.

Hinrichs, Erhard, and Thomas Zastrow. 2007. Novel approaches to computational dialectometry – vector analysis and information theory. In P. Osenova, E. Hinrichs, and J. Nerbonne (Eds.), *International Workshop on Computational Phonology Proceedings*, 37–41, Borovets, Bulgaria, September. INCOMA Ltd.

Johannessen, Janne Bondi, Joel Priestley, Kristin Hagen, Tor Anders Åfarli, and Øystein Alexander Vangsnes. 2009. The Nordic Dialect Corpus – an advanced research tool. In K. Jokinen and E. Bick (Eds.), *NODALIDA 2009 Conference Proceedings*, 73–80.

Joshi, Aravind K., and Bangalore Srinivas. 1994. Disambiguation of parts of speech (or supertags): almost parsing. In *Proceedings of the 15th international conference on Computational Linguistics*, Vol. 1, 154–160, Kyoto, Japan. Association for Computational Linguistics.

Kessler, Brett. 1995. Computational dialectology in Irish Gaelic. In *Proceedings of the European ACL*, 60–67, Dublin.

Kessler, Brett. 2001. *The Significance of Word Lists*. Stanford: CSLI Press.

Kleiweg, Peter, John Nerbonne, and Leonie Bosveld. 2004. Geographic projection of cluster composites. In A. Blackwell, K. Marriott, and A. Shimojima (Eds.), *Diagrams 2004*. Springer-Verlag.

Kondrak, Grzegorz. 2002. *Algorithms for Language Reconstruction*. Doctoral dissertation, University of Toronto.

Kruskal, Joseph. 1964a. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 28:1–27.

Kruskal, Joseph. 1964b. Nonmetric multidimensional scaling: a numerical method. *Psychometrika* 29:115–129.

Leinonen, Therese. 2008. Factor analysis of vowel pronunciation in Swedish dialects. *International Journal of Humanities and Arts Computing* 2(1-2):189–204.

Levenshtein, V I. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii* 163(4):845–848.

Lin, Jinhua. 1991. Divergence measures based on shannon entropy. *IEEE Transactions on Information Theory* 37(1):145–151.

Manning, Christopher, and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Mantel, Nathan. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research* 27:209–220.

Nelson, Gerald, Sean Wallis, and Bas Aarts. 2002. *Exploring Natural Language: working with the British component of the International Corpus of English*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Nerbonne, John, and Wilbert Heeringa. 1997. Measuring dialect distance phonetically. In J. Coleman (Ed.), *Workshop on Computational Phonology*, 11–18, Madrid. Special Interest Group of the Association for Computational Linguistics.

Nerbonne, John, and Wilbert Heeringa. 2001. Computational comparison and classification of dialects. *Dialectologia et Geolinguistica* 69–83.

Nerbonne, John, and Peter Kleiweg. 2003. Lexical distance in LAMSAS. In J. Nerbonne and W. Kretszchmar (Eds.), *Computational Methods in Dialectometry*, Vol. 37, 339–357. Kluwer. Special issue of Computers and the Humanities.

Nerbonne, John, and Wybo Wiersma. 2006. A measure of aggregate syntactic distance. In J. Nerbonne and E. Hinrichs (Eds.), *Linguistic Distances*, 82–90, Sydney, July. International Committee on Computational Linguistics and the Assocation for Computational Linguistics.

Nivre, Joakim, Johan Hall, and Jens Nilsson. 2006a. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of the fifth international conference on language resource and evaluation (LREC2006)*, 2216–2219, Genoa, Italy, May.

Nivre, Joakim, Jens Nilsson, and Johan Hall. 2006b. Talbanken05 : A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*, Genoa, Italy.

Orton, Harold, and Wilfrid J Halliday (Eds.). 1963. *Survey of English Dialects: Basic Materials*. Vol. 1. Leeds: E. J. Arnold and Son.

Petrov, Slav, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Annual Meeting, Association for Computational Linguistics*, Vol. 1, 433–440. Association for Computational Linguistics.

Petrov, Slav, and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Proceeding of the HLT-NAACL*.

Prokić, Jelena. 2007. Identifying linguistic structure in a quantitative analysis of dialect pronunciation. In *Proceedings of the ACL 2007 Student Research Workshop*, 61–66, Prague, Czech Republic, June. Association for Computational Linguistics.

Rosenkvist, Henrik. 2007. The South-Swedish apparent cleft. 239–250.

Sampson, Geoffrey. 2000. A proposal for improving the measurement of parse accuracy. *International Journal of Corpus Linguistics* 5(1):53–68.

Sanders, Nathan C. 2007. Measuring syntactic difference in British English. In *Proceedings of the ACL 2007 Student Research Workshop*, 1–6, Prague, Czech Republic, June. Association for Computational Linguistics.

Sanders, Nathan C. 2008. Comparison of phonological and syntactic distance measures. Unpublished ms., Qualifying Paper, Indiana University.

Sanders, Nathan C., and Steven B. Chin. 2006. Phonological distance measures for cochlear implant users. *Wiener Medizinische Wochenschrift* 156 [Suppl 119]:8.

Sanders, Nathan C., and Steven B. Chin. 2009. Phonological distance measures. *Journal of Quantitative Linguistics* 43(1):96–114.

Séguy, Jean. 1973. La dialectometrie dans l'atlas linguistique de la gascogne. *Revue de linguistique romane* 37:1–24.

Spruit, Marco René. 2008. *Quantitative Perspectives on syntactic variation in Dutch dialects*. PhD thesis, University of Amsterdam, Utrecht. LOT Dissertation Series 174.

Spruit, Marco René, Wilbert Heeringa, and John Nerbonne. 2006. Associations among linguistic levels. In *Comparing Aggregate Syntaxes special session of the Digital Humanities Conference*, Paris.

Ward, Jr., Joe H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58:236–244.

Wiersma, Wybo. 2009. Automatically extracting typical syntactic differences from corpora. BA thesis, University of Groningen.

Curriculum Vitae

QUALIFICATIONS

Dialect Classification/Dialectology Extraction of linguistic, human-interpretable features

Machine learning Statistical (e.g. unsupervised clustering)
Symbolic (e.g. learning and learnability in Optimality theory)

Parsing Adaptation of computer science parsers to learner natural language

Programming Fluent in Python, Haskell, Java/C#, and Scheme
Experience in C++, Perl, Visual Basic, Javascript, F#, and Common Lisp
Experience in prototyping and web programming
Three-time competitor, ACM International Collegiate Programming Contest

EDUCATION

PhD in Linguistics, minor Computer Science Indiana University, 2010

MA Computational Linguistics Indiana University, 2006, GPA 3.98

BA Computer Science minors in French and Spanish, College of the Ozarks, summa cum laude May 2004

EXPERIENCE

August 2009–November 2009 SDET Intern, Microsoft, Bing Search Infrastructure — Wrote software to suggest tests to developers at checkin. Tested the distributed computation system forming the infrastructure of Bing.

August 2005–May 2009 Research Assistant, IU School of Medicine — Investigated linguistic aspects of cochlear implant user development. Developed novel distance measure and compared it to existing measures and human judgments.

May 2005–August 2005 Application Developer, IU Archives of Traditional Music — Wrote Java application to gather cataloging data for videos. Worked with librarians and ethnomusicologists to gather archival metadata requirements.

August 2004–May 2005 Web Programmer, IU Overseas Studies — Maintained database and its web interface. Developed new features on request and maintained office computers.

Summer 2004, Summer 2003 Lead Programmer, Everyware Inc — Implemented web framework for interactive site building. Rewrote existing plug-in architecture and extended existing interface.

Summer 2002 Intern Programmer, SIL International — Programmed import process for translation editor in C++, with a pre-processor in Python.

Spring 2001–Spring 2004 Lab Assistant, College of the Ozarks Foreign Language Lab — Tutored students in Spanish and French. Maintained lab computers and created tracking databases.

PUBLICATIONS

Sanders, N. C. 2009. Phonological distance measures. *Journal of Quantitative Linguistics*, 43:96–114.

Sanders, N. C. 2008. Cluster analysis of phonological distance measures of cochlear implant users. In *Proceedings of the Tenth International Conference on Cochlear Implants and Other Implantable Auditory Technologies*, 113.

Sanders, N. C. 2007. Measuring Syntactic Difference in British English. In *Proceeding of the ACL 2007 Student Research Workshop*, 1–6, Prague, Czech Republic, June.

Sanders, N. C. and Chin, S. B. 2006. Phonological distance measures for cochlear implant users. *Wiener Medizinische Wochenschrift 156*, [Suppl 119] 8

Sanders, N. C. 2004. Compiler error detection techniques applied to natural language processing. In *Proceedings of the Thirty-fifth SIGCSE Technical Symposium on Computer Science Education*. Association for Computing Machinery, 515